NANO LETTERS

2005
Vol. 5, No. 9
1774−1778

DNA-Decorated Carbon Nanotubes for Chemical Sensing

Cristian Staii and Alan T. Johnson, Jr.*

Department of Physics and Astronomy and Laboratory for Research on the Structure of Matter, University of Pennsylvania, Philadelphia, Pennsylvania 19104

Michelle Chen

Department of Material Science and Engineering, University of Pennsylvania, Philadelphia, Pennsylvania 19104

Alan Gelperin

Monell Chemical Senses Center, Philadelphia, Pennsylvania 19104

Received July 1, 2005; Revised Manuscript Received August 4, 2005

ABSTRACT

We demonstrate a new, versatile class of nanoscale chemical sensors based on single-stranded DNA (ss-DNA) as the chemical recognition site and single-walled carbon nanotube field effect transistors (swCN-FETs) as the electronic read-out component. swCN-FETs with a nanoscale coating of ss-DNA respond to gas odors that do not cause a detectable conductivity change in bare devices. Responses of ss-DNA/swCN-FETs differ in sign and magnitude for different gases and can be tuned by choosing the base sequence of the ss-DNA. ss-DNA/swCN-FET sensors detect a variety of odors, with rapid response and recovery times on the scale of seconds. The sensor surface is self-regenerating: samples maintain a constant response with no need for sensor refreshing through at least 50 gas exposure cycles. This remarkable set of attributes makes sensors based on ss-DNA decorated nanotubes very promising for "electronic nose" and "electronic tongue" applications ranging from homeland security to disease diagnosis.

The one-dimensional carbon cage structure of semiconducting single-walled carbon nanotubes (swCNs) makes their physical properties exquisitely sensitive to variations in the surrounding electrostatic environment, whether the swCNs are suspended in liquid or incorporated into field effect transistor (FET) circuits on a substrate.[1−3] Bare and polymer-coated swCNs are reported to be sensitive to various gases,[4−9] but swCNs functionalized with biomolecular complexes hold great promise as molecular probes and sensors[10−13] targeted for chemical species that interact weakly or not at all with unmodified nanotubes. Derivatized swCN-FETs are attractive as electronic-readout molecular sensors due to their high sensitivity, fast response time, and compatibility with dense array fabrication.[3] Derivatized semiconductor nanowires have similar performance advantages, and recent work indicates they are also promising candidates for gas-[14] and liquid-phase sensors.[15,16]

An effective scheme to functionalize swCN-FET sensors should simultaneously achieve robust, reproducible decoration of the swCN with molecular flexibility promising sensitivity to a wide spectrum of analytes. Noncovalent functionalization is required to avoid degrading the high-quality electronic properties of the swCN-FET.

Nucleic acid biopolymers are intriguing candidates for the molecular targeting layer since they can be *engineered*, using directed evolution, for affinity to a wide variety of targets, including small molecules and specific proteins.[17,18] High throughput screening of multiple oligomers was used recently to select films of dye-labeled ss-DNA oligomers that function as gas sensors.[19,20] On exposure to an odor, fluorescence of the intercalated dye changes relative to the level measured when the sample is exposed to clean air. The molecular mechanism of this response is not known, but the response to particular odors was reported to be specific for the base sequence of the oligomer. ss-DNA is also known to have high affinity for swCNs due to a favorable $\pi-\pi-$ stacking interaction.[21]

These findings motivated our exploration of the ss-DNA/swCN-FET hybrid nanostructure as a gas sensor with electronic readout. We focus here on devices consisting of individual nanotubes contacted by electrodes in order to illuminate intrinsic properties of the ss-DNA/swCN system. Sensors based on swCN networks may be easier to manufacture in functional systems, but they introduce complicat-

* Corresponding author. E-mail: cjohnson@physics.upenn.edu.

10.1021/nl051261f CCC: $30.25
Published on Web 08/23/2005

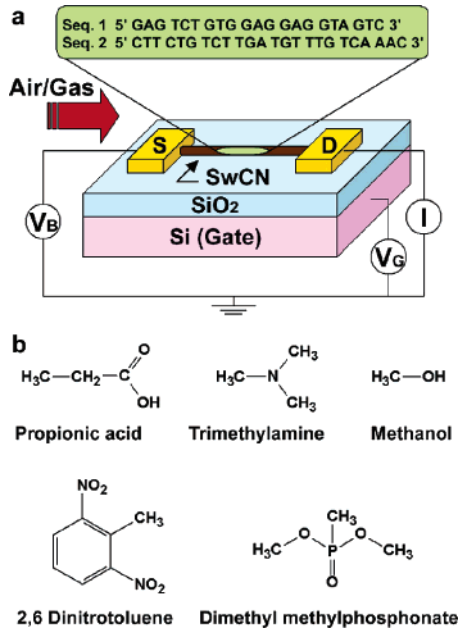


Figure 1. (a) Schematic of the experimental setup. (b) Gases (odors) used in the experiment.

ing, poorly controlled factors, e.g., the presence of both metallic and semiconducting swCNs, and tube−tube cross junctions.

swCNs were grown by catalytic chemical vapor deposition (CVD) on a SiO_2/Si substrate. FET circuits were fabricated with Cr/Au source and drain electrodes patterned using electron beam lithography and the degenerately doped silicon substrate used as a backgate (Figure 1).[22] For each device, source-drain current I was measured as a function of bias voltage V_B and gate voltage V_G under ambient laboratory conditions. Circuits consisting of individual p-type semiconducting nanotubes, where the carriers are positively charged holes, were selected by using only devices that showed a strong decrease in $I(V_G)$ for positive V_G (ON/OFF ratio exceeding 1000).

Two ss-DNA sequences were chosen based on prior work:[19,20]

Sequence 1

5′ GAG TCT GTG GAG GAG GTA GTC 3′

Sequence 2

5′ CTT CTG TCT TGA TGT TTG TCA AAC 3′

Oligonucleotides were obtained from Invitrogen (Carlsbad, CA) and diluted in distilled water to make a stock solution of 658 μg/mL (sequence (Seq) 1) or 728 μg/mL (Seq 2). After odor responses of the bare swCN-FET device were measured, a 500 μm diameter drop of ss-DNA solution was applied to the device for 45 min and then dried in a nitrogen stream. About 25 devices from two different swCN growth runs were selected for detailed analysis and treated with ss-DNA for the experiments described here. Statistical analysis of atomic force microscopy (AFM) images of the same tube

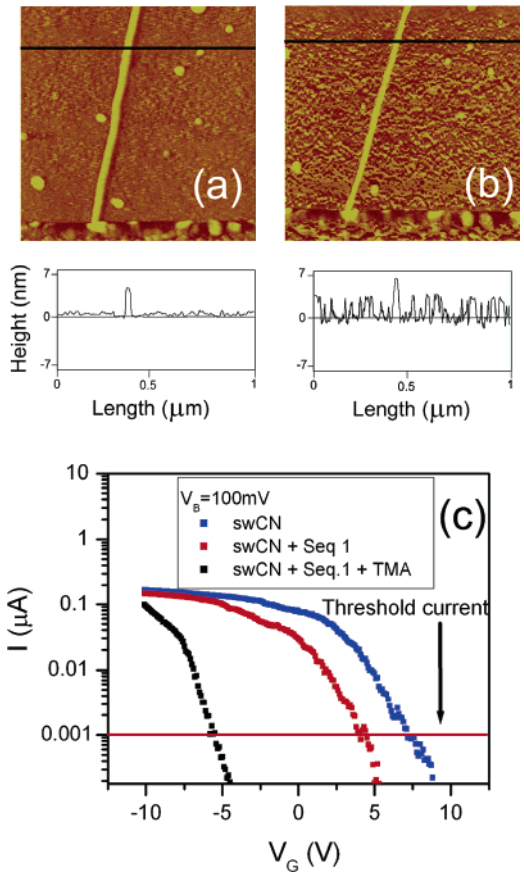


Figure 2. (a, b) AFM images (1 μm × 1 μm, z-range 10 nm) and line scans of the *same* swCN before (a) and after (b) functionalization with ss-DNA. The measured diameter of the bare swCN is 5.4 ± 0.1 nm, while after application of ss-DNA its diameter is 7.2 ± 0.2 nm. The increase in surface roughness in (b) is attributed to nonspecific binding of ss-DNA to the SiO_2 substrate. (c) Current (I) versus backgate voltage (V_G) characteristic of a bare swCN-FET sensor (blue), the same device after functionalization with ss-DNA sequence 1 and exposed to air, the same ss-DNA/swCN-FET exposed to TMA vapor. Source-drain bias voltage is 100 mV.

before and after DNA application showed an increase in the nominal tube diameter from 5.4 ± 0.1 to 7.2 ± 0.2 nm, indicating formation of a nanoscale layer of ss-DNA on the swCN surface (Figure 2a,b). For both of the sequences used, application of ss-DNA caused the threshold value of V_G for measurable conduction to decrease by 3−4 V (Figure 2c). This corresponds to a hole density decrease of roughly 400 μm, assuming a backgate capacitance (25 aF/μm) that is typical for this device geometry.[22] Furthermore, the "ON" state conductivity of the ss-DNA/swCN-FET was ∼10% lower than that of the bare device (Figure 2c), suggesting weak carrier scattering by the molecular coating.

Figure 1b shows the five odors used to characterize the sensor response: methanol, propionic acid, trimethylamine (TMA), dinitrotoluene (DNT), and dimethyl methylphosphonate (DMMP; a simulant for the nerve agent sarin[23]). For DNT, a liquid solution was prepared by dissolving 50 mg/mL of the material in dipropylene glycol.

A reservoir of saturated vapor of each odor was prepared and connected to a peristaltic pump and switching valve array so the flow of room air directed over the device (0.1 mL/s)

Table 1. Measured Responses of Devices to Gaseous Analytes[a]

odor	vapor pressure (Torr)	estimated concn (ppm)	% $\Delta I/I$ bare swCN	% $\Delta I/I$ swCN + Seq 1	% $\Delta I/I$ swCN + Seq 2
water	17.5	700	0 ± 1	0 ± 1	0 ± 1
propionic acid	4	150	0 ± 1	$+17 \pm 2$	$+8 \pm 1$
TMA	500	20000	-9 ± 2	-20 ± 2	-30 ± 2
methanol	100	4000	0 ± 1	-12 ± 2	-20 ± 2
DMMP	0.6	25	0 ± 1	-14 ± 2	-7 ± 2
DNT	1	40	0 ± 1	-14 ± 4	-4 ± 2

[a] Estimated concentration corresponds to 3% of the saturation vapor pressure (www.sciencestuff.com). Each quoted sensor response is based on measurements of 5−10 different devices. Uncertainties are the standard deviation of the mean.

could be electrically diverted to one of the odor reservoirs for a set time (typically 50 s), after which the flow reverted to plain air. The air or air/analyte mixture was directed toward the sample through a 2 ± 0.1 mm diameter nozzle positioned 6 ± 1 mm above the sample surface. For each analyte, we estimate the concentration delivered to the sample to be 3% of the appropriate saturated vapor pressure (see Table 1). The source-drain current (I) through the device was measured as a function of gate voltage V_G for a fixed bias voltage V_B. For each sample (both before and after application of ss-DNA), it was found that $V_G = 0$ V was a region of large transconductance (dI/dV_G), indicating high sensitivity of the swCN-FET to environmental perturbations. Detailed measurements of the odor-induced changes in I as a function of V_G and V_B and sensor response as a function of analyte concentration will be the subject of future work. Here we focus on odor-induced changes in the current measured with $V_B = 100$ mV and $V_G = 0$ V.

In Figure 3a,b, we show responses of two devices to odors, before (blue points) and after (red points) coating with ss-DNA Seq 2. The current response of a bare swCN-FET was less than our experimental sensitivity ($\Delta I/I \sim 1\%$) when exposed to methanol (Figure 3a), propionic acid, DMMP, and DNT (data not shown). After this same device was coated with ss-DNA Seq 2, exposure to methanol gives a 20% decrease in the transport current. We conclude that the ss-DNA layer increases the binding affinity for methanol to the device, thereby increasing the sensor response. Even when a bare swCN sensor responds to a particular gas ($\Delta I/I = -10\%$ on exposure to TMA, Figure 3b), functionalization with ss-DNA enhances the molecular affinity and associated response ($\Delta I/I = -30\%$).

We observe further that different odors elicit different current responses from ss-DNA/swCN-FET sensors. For example, the response to propionic acid of a device with ss-DNA Seq 1 differs in both sign and magnitude from the response to methanol (Figure 3c). The data in Figure 3 also demonstrate a constant sensor response is maintained through multiple odor exposures. As a test of response reproducibility, we exposed a device to 50 cycles of TMA and air exposure (odor and air pulses each 50 s in duration), and the response was maintained to within 5% (data not shown). Device-to-device variation in odor response is also small (see Table 1 and discussion below). This excellent reproducibility for a single device and across devices indicates very favorable prospects for quantitative modeling of individual devices and integrated systems.[24]

Finally we find that the odor response characteristics of ss-DNA/swCN-FET sensors are specific to the *base sequence* of the ss-DNA used (Table 1). The number of distinct ss-DNA 24-mers is extremely large, and they are all expected to bind readily to swCNs through a $\pi-\pi$ stacking interaction. It should be possible to create a large family of sensors with disparate odor response characteristics, an important building block of "electronic nose" and "electronic tongue" systems discussed below.[25]

To explore this possibility, we measured the odor response of ss-DNA/swCN-FET sensors to DNT and DMMP, simulants for explosive vapor and nerve gas, respectively. As seen in Figure 4 and Table 1, ss-DNA functionalized swCN-FETs respond to these two odors while bare devices do not, and the response characteristic is specific to the ss-DNA sequence used to decorate the device. Control experiments were conducted to verify that the ss-DNA/swCN-FET sensor

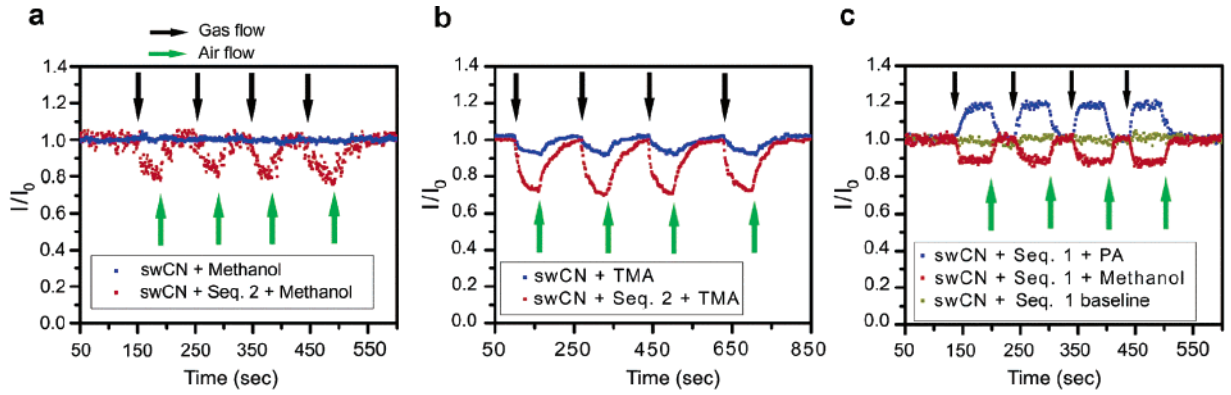


Figure 3. Change in sensor current upon odor exposure. Currents are normalized to I_0, the value when exposed to air (no odor). (a) Bare swCN-FET does not respond to methanol vapor (blue points). The same device coated with ss-DNA sequence 2 (Seq 2) shows clear responses to methanol (red points). (b) A second bare device responds to TMA (blue points), but after application of Seq 2, the response is tripled (red points). (c) The sensor response to propionic acid (blue points) differs in sign and magnitude from the response to methanol (red points). Green data are the current baseline (no odor). $V_B = 100$ mV and $V_G = 0$ V for all data sets.

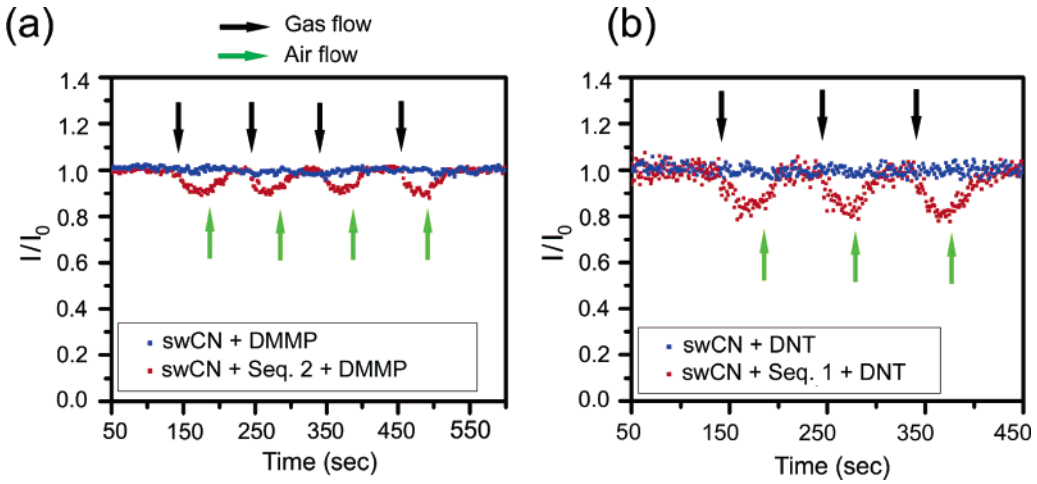


Figure 4. (a) Change in the device current when sarin-simulant DMMP is applied to swCN-FETs before and after ss-DNA functionalization. (b) Sensor response to DNT.

showed no response to dipropylene glycol, the solvent used for DNT, and water vapor, a common background substance. The signal-to-noise levels of the measurements in Figure 4 indicate that detection of concentrations less than 1 ppm should be possible even with these unoptimized devices. The DMMP concentration used in the experiment is estimated to be 25 ppm, and the observed response is distinct but modest ($\Delta I/I \sim -7\%$ for ss-DNA sequence 1 and -14% for ss-DNA sequence 2). Thin-film transistor sensors fabricated from swCN *networks* are reported to respond much more strongly to DMMP ($\Delta I/I \sim -50\%$ at a concentration of 1 ppb[6]). Our experiment indicates this large response is not intrinsic to individual swCNs but may be related to the network connectivity.

We briefly consider the mechanism of molecular detection, recognizing that many important issues remain to be clarified by future experiments. It is known that p-type swCN-FETs can detect analytes through "chemical gating" where a positively (negatively) charged molecular species adsorbs to the nanotube sidewall and locally depletes (enhances) the swCN carrier density leading to a decrease (increase) in current through the FET.[26] This mechanism is consistent with the current decrease and $I(V_G)$ data (Figure 1c) during TMA exposure. Given its pK value of 9.8, TMA should be protonated by residual water (presumably pH $\sim$ 7) associated with the ss-DNA, leading to a rigid shift of the $I(V_G)$ characteristic to the left and a decrease in the sensor current, as observed. The data in Figure 1c correspond to a hole density decrease of $\sim$1200/μm due to TMA exposure. Similarly, propionic acid is expected to donate a proton to residual water, in agreement with the measured increase in sensor current. DMMP, along with other chemical nerve agents, is known to be a strong electron donor,[6] consistent with the observed sensor response ($\Delta I/I < 0$). The detection mechanisms for methanol and DNT are less clear. Simple acid−base considerations suggest methanol is neutral under the experimental conditions, and DNT is expected to be an electron acceptor, inconsistent with the measured current decrease for both odors. More detailed experiments are needed to determine whether these species transfer charge to the swCN in the presence of ss-DNA or if detection occurs through a different mechanism, e.g., a conformational change of the ss-DNA that is transduced into an electrical signal by the swCN-FET.

"Electronic nose" detectors are inspired by olfactory systems in biological organisms that typically utilize a thousand different odor receptors, each responsive to many different odorants, to perform amazing feats of molecular identification and analysis. ss-DNA/swCN-FET gas sensors have a number of properties making them ideal for this application. They are all-electronic sensors with high sensitivity and fast response times (seconds) that compare favorably with well-established and more recently demonstrated sensor families.[27] They offer the advantages of a smaller footprint and simpler implementation than chemicapacitors, and more direct readout with simpler equipment than sensors where odor detection is converted into an optical signal. We demonstrated that the ss-DNA chemical recognition layer is reusable through at least 50 cycles without refreshing or regeneration. It is likely this hybrid nanoscale sensor can be used for liquid phase detection, making it equally appropriate for application in an "electronic tongue" system.[28]

Finally, the intrinsic chemical versatility of ss-DNA, progress in nucleic acid engineering, and use of high-throughput screening may well enable selection of appropriate sequences for detection of a large number of chemical and biological targets. The range of possible targets may be limited by the fact that the ss-DNA chemical recognition component most likely assumes a range of conformations. Future experiments will explore the effectiveness of this sensor class for detection of analytes beyond the small molecules demonstrated here. It has been suggested[29] that an array of 100 sensors with different response characteristics and an appropriate pattern recognition algorithm are sufficient to detect and identify a weak known odor in the face of a strong and variable background. The results presented here represent significant progress toward the realization of a large and diverse sensor array for electronic olfaction and may

bring a practical device within reach when combined with recent progress in fabricating multiplexed arrays of swCN sensors.

Acknowledgment. This work was supported by the Laboratory for Research on the Structure of Matter (NSF DMR00-79909) and by the US Department of Energy, Grant No. DE-FG02-98ER45701 (M.C.). We thank Michael F. Stern and Dr. Karen McAllister for useful discussions and Dr. Douglas R. Strachan for assistance with data acquisition.

References

(1) Kong, J.; Franklin, N. R.; Zhou, C. W.; Chapline, M. G.; Peng, S.; Cho, K. J.; Dai, H. J. *Science* **2000**, *287*, 622−625.
(2) Freitag, M.; Johnson, A. T.; Kalinin, S. V.; Bonnell, D. A. *Phys. Rev. Lett.* **2002**, *89*, art. no.-216801.
(3) Pengfei, Q. F.; Vermesh, O.; Grecu, M.; Javey, A.; Wang, O.; Dai, H. J.; Peng, S.; Cho, K. J. *Nano Lett.* **2003**, *3*, 347−351.
(4) Chopra, S.; McGuire, K.; Gothard, N.; Rao, A. M.; Pham, A. *Appl. Phys. Lett.* **2003**, *83*, 2280−2282.
(5) Li, J.; Lu, Y. J.; Ye, Q.; Cinke, M.; Han, J.; Meyyappan, M. *Nano Lett.* **2003**, *3*, 929−933.
(6) Novak, J. P.; Snow, E. S.; Houser, E. J.; Park, D.; Stepnowski, J. L.; McGill, R. A. *Appl. Phys. Lett.* **2003**, *83*, 4026−4028.
(7) Valentini, L.; Armentano, I.; Kenny, J. M.; Cantalini, C.; Lozzi, L.; Santucci, S. *Appl. Phys. Lett.* **2003**, *82*, 961−963.
(8) Bradley, K.; Gabriel, J. C. P.; Star, A.; Gruner, G. *Appl. Phys. Lett.* **2003**, *83*, 3821−3823.
(9) Snow, E. S.; Perkins, F. K.; Houser, E. J.; Badescu, S. C.; Reinecke, T. L. *Science* **2005**, *307*, 1942−1945.
(10) Wong, S. S.; Joselevich, E.; Woolley, A. T.; Cheung, C. L.; Lieber, C. M. *Nature* **1998**, *394*, 52−55.
(11) Williams, K. A.; Veenhuizen, P. T. M.; de la Torre, B. G.; Eritja, R.; Dekker, C. *Nature* **2002**, *420*, 761−761.
(12) Chen, R. J.; Bangsaruntip, S.; Drouvalakis, K. A.; Kam, N. W. S.; Shim, M.; Li, Y. M.; Kim, W.; Utz, P. J.; Dai, H. J. *P. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 4984−4989.
(13) Barone, P. W.; Baik, S.; Heller, D. A.; Strano, M. S. *Nat. Mater.* **2005**, *4*, 86−92.
(14) Zhang, D.; Liu, Z.; Li, C.; Tang, T.; Liu, X.; Man, S.; Lei, B.; Zhou, C. *Nano Lett.* **2004**, *4*, 1919−1924.
(15) Hahm, J.-i.; Lieber, C. M. *Nano Lett.* **2004**, *4*, 51−54.
(16) Wang, W. U.; Chen, C.; Lin, K.-h.; Fang, Y.; Lieber, C. M. *P. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 3208−3212.
(17) Patel, D. J.; Suri, A. K.; Jiang, F.; Jiang, L. C.; Fan, P.; Kumar, R. A.; Nonin, S. *J. Mol. Biol.* **1997**, *272*, 645−664.
(18) Breaker, R. R. *Nature* **2004**, *432*, 838−845.
(19) White, J. E.; Kauer, J. S. **2004**.
(20) White, J. E.; Williams, L. B.; Atkisson, M. S.; Kauer, J. S. Assoc. Chemoreception Sciences, XXVI Annual Meeting Abstracts **2004**, 32.
(21) Zheng, M.; Jagota, A.; Semke, E. D.; Diner, B. A.; Mclean, R. S.; Lustig, S. R.; Richardson, R. E.; Tassi, N. G. *Nat. Mater.* **2003**, *2*, 338−342.
(22) Radosavljevic, M.; Freitag, M.; Thadani, K. V.; Johnson, A. T. *Nano Lett.* **2002**, *2*, 761−764.
(23) Hopkins, A. R.; Lewis, N. S. *Anal. Chem.* **2001**, *73*, 884−892.
(24) Gao, H.; Kong, Y. *Annu. Rev. Mater. Res.* **2004**, *34*, 123−150.
(25) Gouma, P.; Sberveglieri, G. *MRS Bull.* **2004**, *29*, 697−702.
(26) Kong, J.; Dai, H. J. *J. Phys. Chem. B* **2001**, *105*, 2890−2893.
(27) Katz, H. E. *Electroanalysis* **2004**, *16*, 1837−1142.
(28) Winquist, F.; Krantz-Rulcker, C.; Lundstrom, I. *MRS Bull.* **2004**, *29*, 726−731.
(29) Gelperin, A.; Hopfield, J. J. In *Chemistry of Taste*; Given, P., Ed.; American Chemical Society: Washington, DC, 2002; pp 289−317.

NL051261F

NANO LETTERS

2007
Vol. 7, No. 10
3086−3091

Functionalized Carbon Nanotubes for Detecting Viral Proteins

Yian-Biao Zhang,[†,§] Mandakini Kanungo,[‡,§] Alexander J. Ho,[‡,||] Paul Freimuth,[†] Daniel van der Lelie,[*,†] Michelle Chen,[⊗] Samuel M. Khamis,[⊥] Sujit S. Datta,[⊥] A. T. Charlie Johnson,[*,⊥] James A. Misewich,[*,‡] and Stanislaus S. Wong[*,‡,#]

Biology Department, Brookhaven National Laboratory, Building 463, Upton, New York 11973, Condensed Matter Physics and Materials Science Department, Brookhaven National Laboratory, Building 480, Upton, New York 11973, Biomedical Engineering Department, State University of New York at Stony Brook, Stony Brook, New York 11794, Department of Materials Science and Engineering, University of Pennsylvania, Philadelphia, Pennsylvania 19104, Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, Pennsylvania 19104, and Department of Chemistry, State University of New York at Stony Brook, Stony Brook, New York 11794

Received July 1, 2007; Revised Manuscript Received August 26, 2007

ABSTRACT

We investigated the biocompatibility, specificity, and activity of a ligand−receptor-protein system covalently bound to oxidized single-walled carbon nanotubes (SWNTs) as a model proof-of-concept for employing such SWNTs as biosensors. SWNTs were functionalized under ambient conditions with either the Knob protein domain from adenovirus serotype 12 (Ad 12 Knob) or its human cellular receptor, the CAR protein, via diimide-activated amidation. We confirmed the biological activity of Knob protein immobilized on the nanotube surfaces by using its labeled conjugate antibody and evaluated the activity and specificity of bound CAR on SWNTs, first, in the presence of fluorescently labeled Knob, which interacts specifically with CAR, and second, with a negative control protein, YieF, which is not recognized by biologically active CAR proteins. In addition, current−gate voltage (I–V_g) measurements on a dozen nanotube devices explored the effect of protein binding on the intrinsic electronic properties of the SWNTs, and also demonstrated the devices' high sensitivity in detecting protein activity. All data showed that both Knob and CAR immobilized on SWNT surfaces fully retained their biological activities, suggesting that SWNT−CAR complexes can serve as biosensors for detecting environmental adenoviruses.

In recent years, there has been growing interest in forming viable strategies for the controlled functionalization of single-walled nanotubes (SWNTs) with biological systems.[1−3] Such bio−nano integrated systems, combining the conducting and semiconducting properties of carbon nanotubes with the recognitive and catalytic properties of biomaterials, offer particular promise for developing novel biosensor systems. Specific recognition of target molecules is the essential feature for biological sensing. Accordingly, we have been interested in addressing the following key issues:

1. Do ligand−receptor proteins, bound onto SWNTs, retain their active configuration and conformation so as to remain amenable to biological interactions? We answered this question by demonstrating that functionally active Knob and CAR can be bound onto SWNT surfaces through an amide linkage generated via a diimide reagent.

2. Can this system subsequently be utilized for biological sensing? Our electrical measurements herein demonstrated the applicability of single biofunctionalized carbon nanotubes as field effect transistor (FET) biosensors wherein the biological moiety maintains its activity, proper binding conformation, and biospecificity.

Several different biomolecular systems have been previously affixed to the external surfaces of SWNTs with the goal of creating functional devices. For example, enzyme-coated SWNTs were used as sensors that either modulate their optical properties upon adsorption or alter their conductance upon variations in pH.[4,5] Viruses were employed to assemble SWNTs and other materials into organized networks.[6] Proteins such as ferritin, avidin, bovine serum albumin (BSA), and streptavidin,[7−9] as well as metallopro-

* To whom correspondence should be addressed. E-mail: vdlelied@bnl.gov (D.v.d.L.); cjohnson@physics.upenn.edu (A.T.C.J.); misewich@bnl.gov (J.A.M.); sswong@notes.cc.sunysb.edu (S.S.W.).
[†] Biology Department, Brookhaven National Laboratory.
[‡] Condensed Matter Physics and Materials Science Department, Brookhaven National Laboratory.
[§] These authors contributed equally to this work.
[||] Biomedical Engineering Department, State University of New York at Stony Brook.
[⊗] Department of Materials Science and Engineering, University of Pennsylvania.
[⊥] Department of Physics and Astronomy, University of Pennsylvania.
[#] Department of Chemistry, State University of New York at Stony Brook.

10.1021/nl071572l CCC: $37.00
Published on Web 09/26/2007

teins and enzymes[10] have been noncovalently bound onto SWNT surfaces so as to generate highly specific electronic biosensors. Other groups have coated peptides[11] with selective affinity for SWNTs onto their surfaces while different laboratories have bound proteins[12,13] such as either ferritin or BSA onto oxidized SWNTs via an amide linkage in aqueous solution. In these latter studies, the biological activities of the attached moieties were confirmed but the electrical activity of the functionalized nanotubes was not measured.

Although this is a relatively unexplored area of research, what is abundantly evident is that the covalent functionalization of biologically active ligand−receptor proteins onto single SWNTs and SWNT bundles clearly affords a viable strategy toward developing specific, quantitative biosensors. Precedence for such a strategy lies in a few unrelated studies. For instance, complementary detection of prostate-specific antigen was demonstrated by using In_2O_3 nanowire and SWNT devices.[14] SWNT-FET-based biosensors composed of either DNA aptamers[15] or single-stranded DNA[16] as molecular recognition elements were also reported, although most of this DNA work relied on noncovalent interactions with the SWNTs.

In the present study, we demonstrate a simple, fast-response, highly sensitive, real-time biosensor composed of a ligand−receptor protein complex covalently attached by a diimide linker to oxidized SWNTs via a mild, ambient, straightforward, and economical protocol. That is, we not only retained the intrinsic biological activity and specificity of the attached complex but also conserved the highly favorable electronic properties of SWNTs in these biofunctionalized single-tube devices. The proteins we used were the adenovirus protein, Ad12 Knob, and its complementary human "Coxsackie virus and adenovirus receptor", CAR.

Adenoviruses are one of many subclasses of viruses that cause infections such as the common cold and mild ailments of the upper respiratory and gastrointestinal tracts. Unlike viruses such as HIV, Ebola, and poliovirus, adenoviruses do not use either envelope proteins or capsid domains to infect cells. Rather, infection is initiated by the formation of a high affinity complex between the Knob trimer and its complementary adenovirus CAR receptor present in human cells. Upon binding CAR, the Knob-coated virus replicates within the cell nucleus, triggering infections.[17,18] Currently, adenoviruses are the leading candidates as vectors for gene therapy.[19] In our work, we used 6 mg/mL of purified Knob and 2.5 mg/mL of CAR protein, as verified by using a BCA assay kit.

Raw HiPco (high-pressure carbon monoxide decomposition process) SWNT bundles as well as individual SWNTs (prepared on surfaces by in situ catalytic chemical vapor deposition) were purified and air oxidized by using a modification of the gasification−dissolution method described earlier by Chiang et al.[20] This process generates surface functionalities on the nanotubes, particularly carboxylic acids at their ends and sidewalls. Air-oxidized SWNTs were then suspended in a 50 mM phosphate buffer (pH 8) solution at a concentration of 1 mg/mL. Proteins were attached to the processed SWNTs via a two-step process of carbodiimide (EDAC)-mediated activation previously described.[14] We

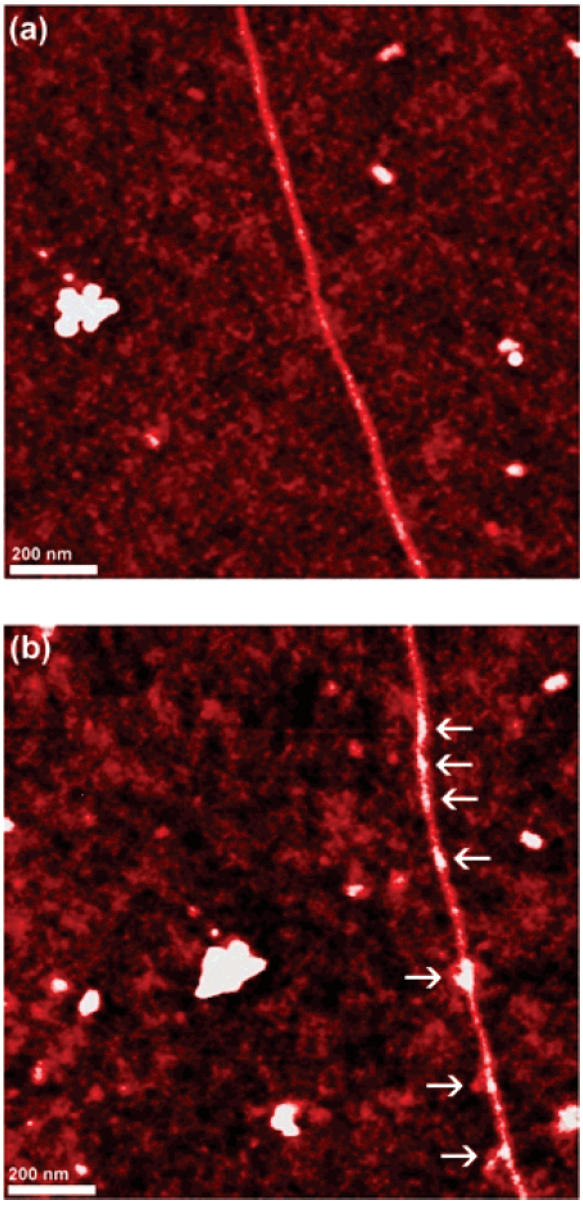


Figure 1. (a) AFM height image of single-walled carbon nanotubes on a Si/SiO_2 substrate after oxidation. The *z*-axis color scale is 10 nm. (b) AFM height image of the same nanotube after the attachment of the CAR protein and subsequent exposure to Knob protein. Arrows indicate seven main sites along the nanotube's length where the CAR + Knob complex is bound to the nanotube's surface. The vertical *z*-axis color scale is 10 nm. The images were low-pass filtered for clarity.

confirmed that the proteins had, indeed, bound to the SWNTs by atomic force microscopy (AFM) height analysis. Further experimental details, including protein labeling, are given in the Supporting Information (Figures S1−S3).

We obtained AFM measurements before and after protein attachment on four samples. Representative AFM height images and statistical analysis of selected cross sections (shown in Figure 1) conclusively showed that this functionalization procedure attaches protein complexes (CAR + Knob) along the length of the SWNT (on average, ∼1 μm for individual tubes). The observed protein density in Figure 1 was approximately 1 per 200 nm, but in other samples, it approached the limit of the AFM resolution (∼1 per 20 nm).

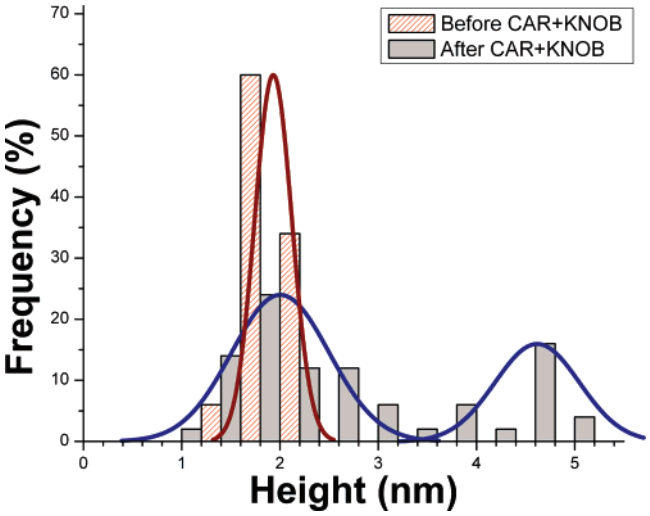


Figure 2. Histogram of SWNT diameters based on 50 equally spaced height sections from each AFM image with Gaussian fits. Before applying any proteins, the data (red, striped bars and red fit) show a SWNT diameter of 1.9 ± 0.2 nm. After applying CAR + Knob proteins, the distribution (gray bars and blue fits) exhibits *two* peaks, corresponding to heights of 2.0 ± 0.5 nm and 4.4 ± 4 nm, ascribed to regions of the nanotube without and with attached protein complexes, respectively.

Height analysis (Figure 2) demonstrated that the diameter of an individual SWNT bundle after oxidation was 1.9 ± 0.2 nm and that the additional height observed, associated with the attached protein complex (CAR + Knob), was 2.5 ± 0.2 nm. AFM measurements on five other samples exposed to CAR alone yielded a height increase of 0.5 nm (data not shown) so that by extension, the intrinsic height increase due to Knob itself was approximately 2 nm. These values are somewhat smaller than the accepted molecular sizes of these proteins, consistent with our expectation that (i) pressure from the AFM tip may have distorted the protein and that (ii) proteins attached to the mostly hydrophobic nanotube surface in air can be slightly different from their intrinsic morphology in aqueous solution. As demonstrated later and further corroborated in the Supporting Information (data on SWNT bundles, Figure S5), CAR proteins effectively attached to individual SWNTs as a single layer and retained their critical molecular recognition functionality.

After confirming the formation of our protein−SWNT constructs, we explored their interaction with both labeled complementary (Ad 12 Knob) and noncomplementary (YieF) proteins, thereby enabling us to assess the activity and specificity of the bound, attached proteins. Both Ad 12 Knob and YieF were labeled by using Alexa Fluor (Molecular Probes). All optical and fluorescence images of the labeled proteins were recorded by using a Zeiss Axiovert 200 fluorescence microscope.

The biological activity of bound Ad 12 Knob was investigated by targeting rhodamine-labeled anti-Knob antibodies to the oxidized SWNT−Knob constructs. These antibodies were purified by using an affinity column of immobilized Ad 12 Knob, and hence, we targeted only those Knob proteins folded in specifically active conformations. Nonspecific binding of rhodamine-labeled anti-Ad 12 Knob protein attached to carbon nanotubes was prevented by blocking this reaction with 4% milk, which contains a number of unrelated, nonspecific proteins in high concentrations. Figure 3a shows the optical image (left) and the corresponding fluorescence image (right) of fluorescently labeled anti-Knob antibodies targeting Ad 12 Knob protein bound to the carbon nanotubes. The fluorescence of the functionalized carbon nanotubes confirms that Ad 12 Knob bound to SWNTs indeed retains its biologically active conformation. As control experiments, labeled anti-Knob antibodies lacking protein were targeted to SWNTs either in the presence (blocked) or absence of milk (not blocked). In the latter case, we observed that the sample fluoresced, suggesting that the labeled antibodies bound to the SWNTs. Conversely, blocked SWNTs exhibited little or no fluorescence (Supporting Information Figure S4), demonstrating the efficacy of milk as a blocking agent. Hence, it is evident that labeled anti-Knob antibodies could specifically target bound Knob proteins on carbon nanotube surfaces.

By analogy, to investigate the biological activity and specificity of bound CAR, we used fluorescently labeled Knob, which shows high specific binding to CAR. In a separate control experiment, we attached YieF, an unrelated 22.4 kDa protein isolated from *E. coli* bacteria, to the SWNTs; YieF is nonspecific for CAR. Hence, the presence of bound CAR proteins in their biologically active conformation will show specificity to Knob proteins. SWNT−CAR constructs were blocked by 4% milk to prevent nonspecific binding of the labeled proteins to the nanotubes. Figure 3b shows an optical image (left) and the corresponding fluorescent image (right) of SWNT−CAR constructs targeted by fluorescently labeled Ad 12 Knob. The sample fluoresces, indicating that Knob is bound to the CAR proteins. On the other hand, after replacing the labeled Ad 12 Knob with labeled YieF, the samples did not fluoresce (Figure 3c). This observation afforded the following evidence: (1) CAR is bound to the carbon nanotubes, (2) bound CAR is biologically active, and (3) CAR-functionalized nanotubes will specifically bind to Ad 12 Knob. Thus, this construct provides us with the basis for a biological sensor to detect the presence of the Ad 12 Knob viral protein.

We measured current−gate voltage ($I-V_g$) data on a dozen nanotube devices to explore the effect of the attachment process and of protein binding on the SWNTs' electronic properties. Typical data are displayed in Figure 4. Nanotube FET devices were of high quality; they consisted of individual SWNTs with ON/OFF ratios exceeding 1000 and possessed on-state resistance values of 100−500 kΩ. Findings discussed below were reproduced in all the devices, although there was some scatter, as noted, in individual responses.

All devices showed a hysteretic $I-V_g$ response, as is typical of nanotube FETs on untreated silica substrates. This response results from charge injection from the nanotube into nearby regions due to the substantial electric field (~10 V/nm) existing at the SWNT surface associated with a large gate voltage (V_g).[21,22] The electric field of this injected

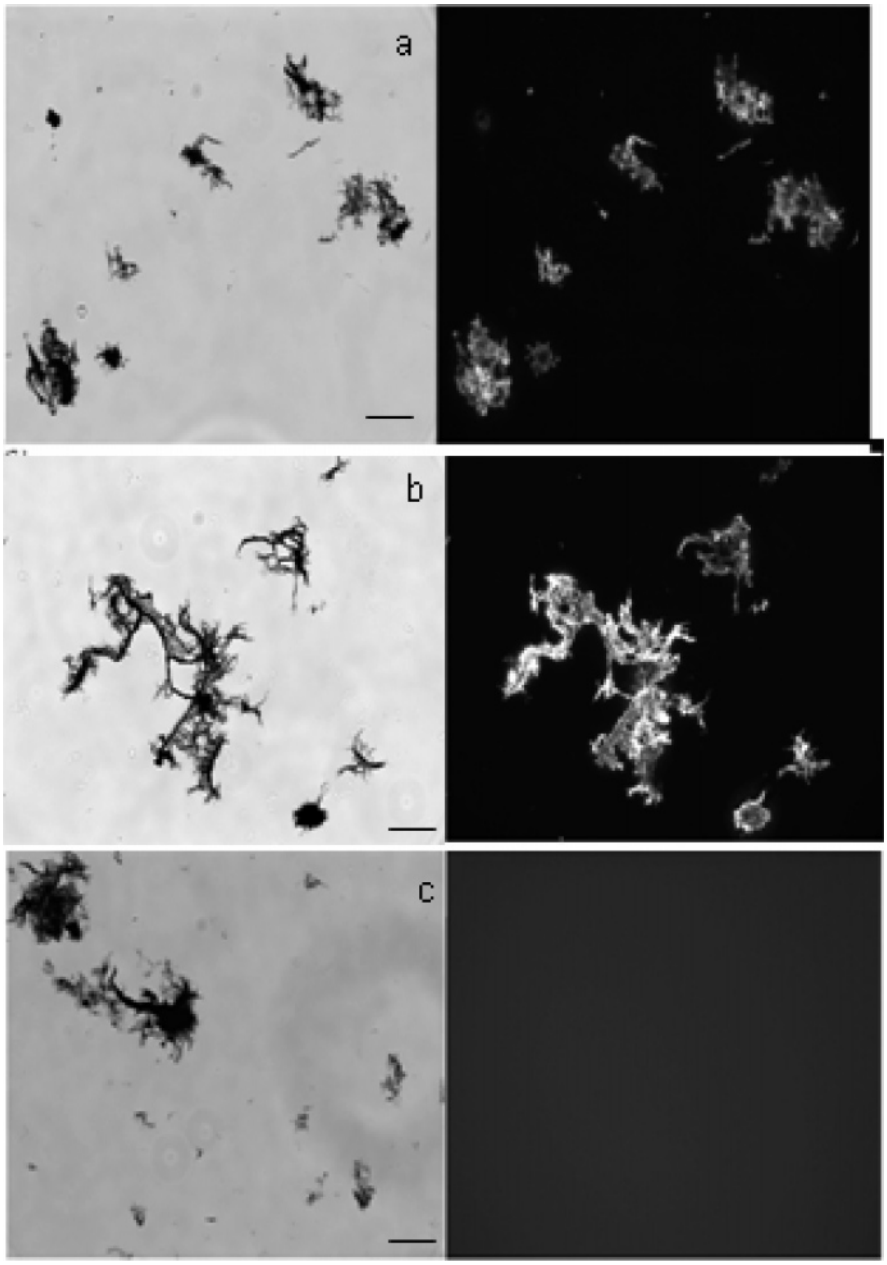


Figure 3. (a) Optical (left) and corresponding fluorescence (right) images of rhodamine-labeled anti-Knob antibodies targeting Ad12 Knob functionalized air-oxidized SWNTs. (b) and (c) show, respectively, the optical and corresponding fluorescence images of fluorescently labeled Ad 12 Knob and YieF targeting CAR-functionalized air-oxidized SWNTs. The sample in (b) fluoresces because the Ad 12 Knob is bound to CAR. On the other hand, there is no observable fluorescence in the sample in (c), where labeled YieF targets functionalized CAR air-oxidized SWNTs. All samples were blocked with milk to prevent the nonspecific binding of proteins onto the SWNT surfaces. The scale bar is 2.5 μm.

charge and of other charge traps near the SWNT is partially screened when the FET is in its ON state, while almost no screening occurs when the FET is in the OFF state. Hence, in the following discussion, we assume that the leftmost (ON-to-OFF) transition of the $I-V_g$ characteristic is more reproducible than the OFF-to-ON transition in the presence of unavoidable charge switching and is, therefore, more amenable to physical interpretation.

The oxidation process typically either increased the ON state current of the device (by 10−25%) or left it unchanged (Figure 4). We concluded that mild oxidation created a low density of defect sites that did not degrade electron transport in the device; we attribute the small increase in the ON state current to contact annealing. Oxidation also generated a reproducible increase of 0.5−3 V in the ON−OFF threshold voltage, consistent with the notion that defect sites created by oxidation are functionalized with oxygenated moieties (such as predominantly carboxyl groups) that become deprotonated in the presence of adsorbed water. This change leaves the groups negatively charged, so that a more positive value of V_g is needed to turn the FET OFF. Assuming a typical backgate capacitance of 25 aF/μm for this geometry, this shift in V_g corresponds to an increase in the carrier density of 80−400 holes/μm.

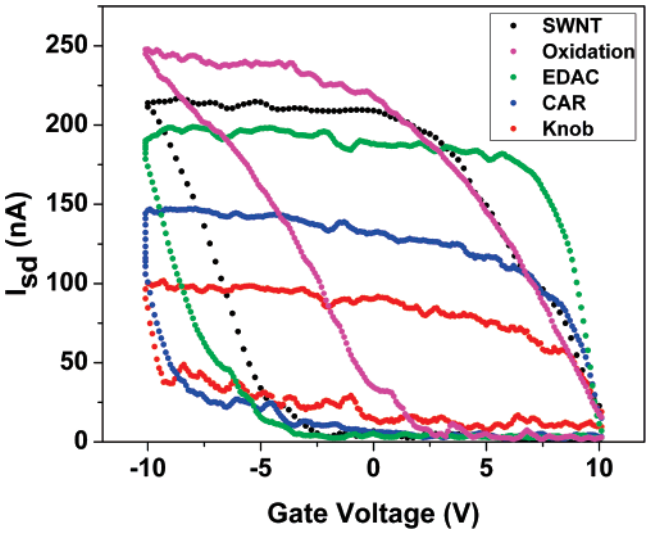


Figure 4. Measured source−drain current as a function of gate voltage for a SWNT FET demonstrates that covalently bound CAR protein retains its molecular recognition functionality. Data are shown for as-grown SWNT (black), SWNTs after oxidation (purple), after exposure to EDAC/NHS (green), upon CAR attachment (blue), and after exposure to the complementary Knob protein (red). The data indicate that Knob specifically binds to the CAR, leading to a significant decrease (~33%) in the ON-state current of the FET. The bias voltage is 100 mV for all measurements.

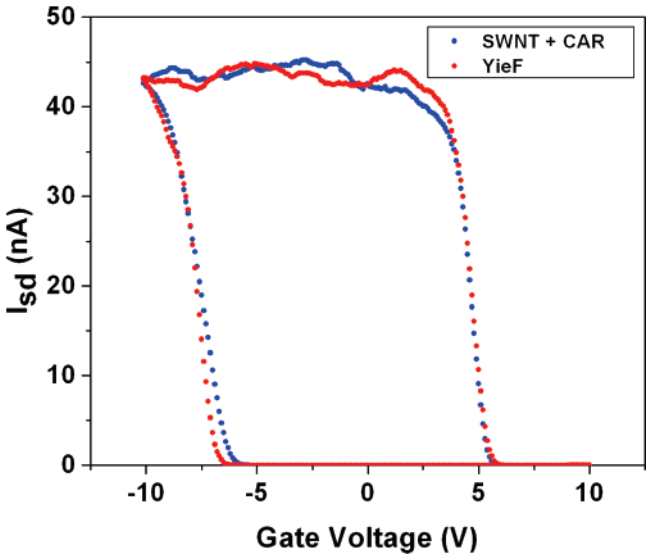


Figure 5. Measured source−drain current vs gate voltage for a SWNT sensor functionalized with CAR protein (blue data) shows no response when exposed to the nonspecific YieF protein (red).

Subsequently, incubating the device in EDAC/NHS solution engendered a negative shift of the threshold voltage to its original value or to an even more negative voltage, in agreement with the expectation that the proton had been replaced by a stable active ester.[12] CAR protein attachment led to a 1−2 V decrease in the ON−OFF threshold voltage and a corresponding 20−40% decrease in the ON-state current. These observations are consistent with the FET experiencing a positive charge and enhanced carrier scattering due to the presence of the protein, as other groups have proposed.[7,23,24] Finally, exposing the CAR−SWNT hybrid to the complementary Knob protein further suppresses the ON state current; the molecular recognition event was thus fully detectable in this system (Figure 4). To quantify the extent of protein binding, we can reasonably assume that the nanotube is 1 μm long, with a corresponding protein density of 1 per 20 nm. Hence, on average, 50 proteins coat the nanotube and the corresponding decrease in current we observed is approximately 1 nA/protein, a sizable value enabling the detection of single protein molecules. Because the applied voltage is 100 mV, the resistance of CAR and of Knob bound to CAR is approximately 667 kΩ and 1 MΩ, respectively, implying a resistance of about 5 kΩ/protein. Using the Landauer formulation in the incoherent transport regime, we thereby obtain a reflection coefficient of approximately 40% per protein, implying that the protein complex is closely bound to the nanotube surface.

In a separate control experiment (Figure 5), CAR-functionalized devices showed no evident change in $I-V_g$ response, as expected, after exposure to (noncomplementary) YieF, implying that the in vivo chemical specificity of the CAR protein is retained even when it is immobilized on the SWNT surface. In another experiment, we noted that the electrical profile of SWNTs, which had been noncovalently functionalized with CAR proteins, reverted to its original signal upon extended washing with phosphate buffer and water; these data highlighted the importance of covalent protein binding in our experiments and implied that weakly bound, physically absorbed proteins were lost upon washing of the SWNTs (data not shown).

The present study provides proof-of-concept for developing a simple, efficient, sensitive, fast-response, and real-time miniaturized nanotube FET biosensor for detecting the Ad 12 Knob virus using CAR−Knob specificity. Moreover, this methodology can be extended to uncover the presence of serotype12 and all other possible CAR-binding adenoviruses (about 30 serotypes, including Ad2 and Ad5), as well as subgroup B Coxsackie viruses. This is the first evidence of straightforward, ambient covalent immobilization of a viral ligand−receptor-protein system onto individual SWNTs and SWNT bundles and of our subsequent confirmation of the bound proteins' retention of biological activity and specificity, as revealed by systematic electrical measurements. Our future goal will be to develop a single-molecule biosensor based on the conductivity change of a single SWNT by adding a discrete CAR domain to the nanotube.

Acknowledgment. Y.Z., D.v.d.L., J.M., and S.S.W. acknowledge support of this work through the U.S. Department of Energy Office of Basic Energy Sciences under contract DE-AC02-98CH10886. S.S.W. also thanks the National Science Foundation (CAREER DMR-0348239) as well as the Alfred P. Sloan Foundation for supplies and financial support. Work conducted in the laboratory of A.T.C.J. was supported by the JSTO DTRA as well as the Army Research Office grant no. W911NF-06-1-0462; research at UPenn was also partially supported by the Nano/Bio Interface Center through the National Science Foundation under contract NSEC DMR-0425780. We thank A. Woodhead for helpful comments.

Supporting Information Available: Expression and purification of AD 12 Knob; CAR protein purification; YieF purification; fluorescent labeling of proteins; preparation of SWNT−protein hybrids; microscopy characterization of SWNTs and of SWNT−protein hybrids; attachment of labeled proteins onto SWNT−protein hybrids; growth, fabrication, and electrical measurements of SWNTs; additional AFM height measurements on SWNT bundles.

References

(1) Katz, E.; Willner, I. *ChemPhysChem* **2004**, *5*, 1084−1104.
(2) Bianco, A.; Prato, M. *Adv. Mater.* **2003**, *15*, 1765−1768.
(3) Wong, S. S.; Joselevich, E.; Woolley, A. T.; Cheung, C. L.; Lieber, C. M. *Nature* **1998**, *394*, 52−55.
(4) Baron, P. W.; Baik, S.; Heller, D. A.; Strano, M. S. *Nat. Mater.* **2005**, *4*, 86−92.
(5) Besteman, K.; Lee, J.-O.; Wiertz, F. G. M.; Heering, H. A.; Dekker, C. *Nano Lett.* **2003**, *3*, 727−730.
(6) Portney, N. G.; Singh, K.; Chaudhary, S.; Destito, G.; Schneemann, A.; Manchester, M.; Ozkan, M. *Langmuir* **2005**, *21*, 2098−2103.
(7) Chen, R. J.; Bangsarnutip, S.; Drouvalakis, K. A.; Kam, N. W. S.; Shim, M.; Kim, W.; Utz, P. J.; Dai, H. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 4984−4989.
(8) Shim, M.; Kam, N. W. S.; Chen, R. J.; Li, Y.; Dai, H. *Nano Lett.* **2002**, *2*, 285−288.
(9) Chen, R. J.; Zhang, Y.; Wang, D.; Dai, H. *J. Am. Chem. Soc.* **2001**, *123*, 3838−3839.
(10) Azamian, B. R.; Davis, J. J.; Coleman, K. S.; Bagshaw, C. B.; Green, M. L. H. *J. Am. Chem. Soc.* **2002**, *124*, 12664−12665.
(11) Wang, S.; Humphreys, E. S.; Chung, S.-Y.; Delduco, D. F.; Lustig, S. R.; Wang, H.; Parker, K. N.; Rizzo, N. W.; Subramoney, S.; Chiang, Y.-M.; Jagota, A. *Nat. Mater.* **2003**, *2*, 196−200.
(12) Jiang, K.; Schadler, L. S.; Seigel, R. W.; Zhang, X.; Zhang, H.; Terrones, M. *J. Mater. Chem.* **2004**, *14*, 37−39.
(13) Huang, W.; Taylor, S.; Fu, K.; Lin, Y.; Zhang, D.; Hanks, T. W.; Rao, A. M.; Sun, Y.-P. *Nano Lett.* **2002**, *2*, 311−314.
(14) Li, C.; Curreli, M.; Lin, H.; Lei, B.; Ishikawa, F. N.; Datar, R.; Cote, R. J.; Thomson, M. E.; Zhou, C. *J. Am. Chem. Soc.* **2005**, *127*, 12484−12485.
(15) So, H.-M.; Won, K.; Kim, Y. H.; Ryu, B. H.; Na, P. S.; Kim, H.; Lee, J.-O. *J. Am. Chem. Soc.* **2005**, *127*, 11906−11907.
(16) Staii, C.; Johnson, A. T., Jr.; Chen, M.; Gelperin, A. *Nano Lett.* **2005**, *5*, 1774−1778.
(17) Freimuth, P.; Springer, K.; Berard, C.; Hainfeld, J.; Bewley, M.; Flanagan, J. M. *J. Virol.* **1999**, *73*, 1392−1398.
(18) Bewley, M. C.; Springer, K.; Zhang, Y.-B.; Freimuth, P.; Flanagan, J. M. *Science* **1999**, *286*, 1579−1583.
(19) Nabel, G. J. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 324−326.
(20) Chiang, I. W.; Brinson, B. E.; Huang, A. Y.; Willis, P. A.; Bronikowski, M. J.; Margrave, J. L.; Smalley, R. E.; Hauge, R. H. *J. Phys. Chem. B* **2001**, *105*, 8297−8301.
(21) Radosavljevic, M.; Freitag, M.; Thadani, K. V.; Johnson, A. T. *Nano Lett.* **2002**, *2*, 761−764.
(22) Führer, M. S.; Kim, B. M.; Durkop, T.; Brintlinger, T. *Nano Lett.* **2002**, *2*, 755−759.
(23) Chen, R. J.; Choi, H. C.; Bansaruntip, S.; Yenilmez, E.; Tang, X. W.; Wang, Q.; Chang, Y. L.; Dai, H. *J. Am. Chem. Soc.* **2004**, *126*, 1563−1568.
(24) Star, A.; Gabriel, J. C. P.; Bradley, K.; Grüner, G. *Nano Lett.* **2003**, *3*, 459−463.

NL071572L

Simultaneous Quality and Reliability Optimization for Microengines Subject to Degradation

Hao Peng, Qianmei Feng, and David W. Coit, *Member, IEEE*

***Abstract*—Micro-Electro-Mechanical Systems (MEMS) represent an exciting new technology, but to achieve more widespread usage and wider adoption within more industrial applications, they must be highly reliable, and manufactured to stringent quality standards. Many challenging manufacturing issues are of concern during the fabrication of MEMS, such as precise dimensional inspection, reliability modeling, burn-in scheduling, avoiding stiction, and maintenance strategies. However, only limited mathematical tools for improving MEMS reliability, quality, and productivity are currently available. This paper proposes a mathematical model to jointly determine inspection & preventive replacement policies for surface-micromachined microengines subject to wear degradation, which is a major failure mechanism for certain MEMS devices. The optimal specification limits for inspection, and the replacement interval are determined by simultaneously optimizing MEMS quality and reliability. The proposed model can be used as a tool for decision-makers in MEMS manufacturing to make sound economical and operational decisions on reliability, quality, and productivity. While illustrated considering one specific microengine design, the proposed model can be applied to a broader range of MEMS devices that experience wear degradation between rubbing surfaces.**

***Index Terms*—Burn-in, MEMS reliability, preventive replacement, quality and reliability optimization, specification limits, wear degradation.**

ACRONYM[1]

pdf	probability density function
cdf	cumulative distribution function
MEMS	Micro-Electro-Mechanical Systems
NDE	Non-Destructive Evaluation
SQP	Sequential Quadratic Programming

NOTATION

t	Number of revolutions to failure
$X(t)$, $X(t;\boldsymbol{\beta})$	Wear volume of material at t (sometimes expressed as a function of model coefficients $\boldsymbol{\beta}$)
H	Critical wear volume or failure threshold
r	Radius of the pin joint
c	Model parameter proportional to the wear coefficient, and inversely proportional to the hardness of the material
F	Force between the contacting surfaces
t_0	Burn-in time, or number of revolutions to burn-in
η	Upper specification limit
$L(X(t_0))$	Quality loss function after burn-in
$C_Q(\eta, t_0)$	Expected quality loss after burn-in
s	Scrap cost per unit
$C_S(\eta, t_0)$	Expected scrap cost
C_I	Inspection cost per unit
QC	Expected quality-related cost
f_c	Cost of failure per unit
FC	Expected failure cost
RC	Replacement cost
τ	Replacement time
B_τ	Upper bound of the replacement interval
$\phi(\cdot)$	pdf of a standard normally distributed variable
$\Phi(\cdot)$	cdf of a standard normally distributed variable

Manuscript received September 16, 2007; revised August 01, 2008; accepted September 13, 2008. First published February 10, 2009; current version published March 04, 2009. The work of H. Peng and Q. Feng was supported by the Grants to Enhance and Advance Research (GEAR) Program at the University of Houston. Associate Editor: L. Cui.

H. Peng and Q. Feng are with the Department of Industrial Engineering, University of Houston, Houston, TX 77204 USA (e-mail: hao_png@yahoo.com; qmfeng@uh.edu).

D. W. Coit is with the Department of Industrial and Systems Engineering, Rutgers University, Piscataway, NJ 08854 USA (e-mail: coit@rutgers.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TR.2008.2011672

[1]The singular and plural of an acronym are always spelled the same.

I. INTRODUCTION

TO ACHIEVE WIDESPREAD usage, Micro-Electro-Mechanical Systems (MEMS) must be highly reliable, and manufactured to stringent quality standards. MEMS technology shows great promise for many critical applications in aerospace, biological/medical, nuclear, and weapons areas. In addition to new applications enabled by MEMS technology, existing applications are enhanced by miniaturized, low-cost, high-performance, and "smart" MEMS technology. MEMS devices have been effectively used in many commercial products, such as accelerometers in automotive airbag deployment systems [13], and inkjet print heads [29]. With more widespread commercialization of MEMS products, many challenging manufacturing/fabrication issues are of concern including precise dimensional control and inspection, reliability testing and modeling, avoiding stiction, and maintenance strategies. These reliability, quality, and productivity issues are dominant factors that impact the process of MEMS moving from the laboratory into the mainstream market. Decision-makers in MEMS manufacturing need tools to optimize these operational decisions.

However, such mathematical tools to improve MEMS reliability, quality, and productivity are currently lacking.

This study proposes a mathematical model to jointly determine inspection and preventive replacement policies for the surface-micromachined microengines subject to wear degradation, which is a major failure mechanism in MEMS devices [24], [25]. The optimal specification limits for inspection, and the replacement interval are determined by optimizing MEMS quality and reliability simultaneously.

A. Failure Analysis of MEMS

Reliability, and quality are important factors for MEMS to evolve from prototypes to commercialization. These issues for MEMS are complicated due to both electronic and mechanical parts, and their interactions [14]. Sufficient understanding of failure mechanisms is required to improve reliability and quality of MEMS devices in critical applications. However, knowledge is still somewhat limited on MEMS failures, and failure causes, at least in the public domain. According to their operational interactions, MEMS devices can be categorized into four classes [22], [23], [30]: Class I devices have no freely moving parts, but may have parts which stretch, compress, or bend, such as accelerometers, pressure sensors, or strain gauges; Class II devices have moving parts without rubbing or contacting surfaces, such as gyros, resonators, and filters; Class III devices have moving parts with contacting surfaces, such as relays, and valve pumps; and Class IV devices have moving parts with rubbing, contacting surfaces, such as shutters, scanners, and optical switches.

Designed with no rubbing surfaces, the first three classes of MEMS devices can achieve a high level of reliability if they are properly manufactured, and packaged. For Class IV MEMS devices, in which rubbing surfaces are unavoidable, failure analysis, and reliability assessment must be performed to further advance the growing commercialization of MEMS. Failure modes, and reliability models of Class IV MEMS were investigated by researchers at Sandia National Laboratories, a leader of MEMS technology [24], [25]. They conducted their research by performing many experiments on a reliability testing infrastructure. The MEMS device used in the reliability testing is a surface-micromachined microengine, developed at Sandia. As shown in Fig. 1, the microengine consists of orthogonal linear comb drive actuators that are mechanically connected to a rotating gear. The linear displacement of the comb drives is transformed to the gear via a pin joint. The gear rotates about a hub that is anchored to the substrate [26].

The dominant failure mechanism is identified as visible wear on rubbing surfaces, which often results in either seized microengines, or microengines with broken pin joints [21], [30]. Wear can be defined as the removal of material from solid surfaces as a result of mechanical actions. Wear degradation is a very complex phenomenon, involving both the mechanical and chemical properties of the bodies in contact, and also the pressure and interfacial velocity under which the bodies make contact.

B. Literature Review on Degradation Processes

Wear processes are degrading phenomena that have been studied in electronics, and other engineering fields. Lu & Meeker [16], and Meeker *et al.* [19] developed general statistical models to estimate the time-to-failure distribution from degradation measures. A general model, and several examples were provided for the degradation path model. Bae & Kvam [2] introduced a log-linear degradation model with an unknown change point to characterize nonlinear degradation paths representing incomplete burn-in during the manufacturing process of plasma display panels. Kharoufeh [11] derived the explicit probability distribution of the random failure time for single-unit systems that deteriorate continuously and additively due to the influence of a random environment modeled as a general, finite-state Markov process. Kharoufeh & Cox [12] presented a degradation-based procedure for the estimation of full, and residue lifetime distribution for single-unit systems using real sensor data. Boulanger & Escobar [3], Tseng *et al.* [27], Yu & Chiao [31], and Joseph & Yu [10] used experimental design to improve reliability for degradation processes.

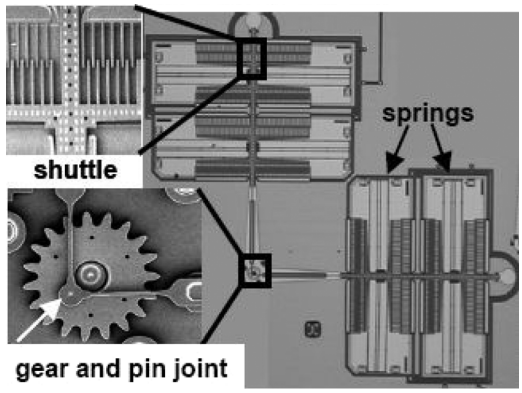


Fig. 1. Scanning electron microscopy image of a microengine [26] (courtesy of SPIE).

For degradation processes, preventive replacement (PR), or preventive maintenance (PM) is often considered as a policy to reduce the number of failures. Many different PR or PM approaches for degrading systems have been studied in the literature. Grall *et al.* [8] developed a maintenance cost model for determining the optimal inspection schedule and replacement threshold for a single unit degrading system. To reduce the uncertainty in cost estimates, Liao *et al.* [15] proposed a condition-based maintenance model for a continuous degradation process by considering imperfect maintenance, and a short-run availability constraint. Lu *et al.* [17], and Lu *et al.* [18] presented a preventive condition-based maintenance approach based on monitoring, modeling, and predicting a system's deterioration. Drapella & Kosznik [4] developed a model to seek for equilibrium of burn-in, and preventive replacement periods. Jiang & Jardine [9] examined the effectiveness of a jointly applied burn-in and preventive replacement policy for situations where the failure time follows a mixture distribution.

C. Research Objectives and Contribution

Although many different preventive replacement approaches for degrading systems have been studied, a comprehensive approach to jointly determine the inspection and preventive replacement policies is not available in the literature. For the microengine wear degradation, this paper proposes a mathematical model to jointly determine the parameters for *inspection,*

and *preventive replacement* policies. For systems with degradation characteristics, manufacturing decisions should be determined by taking into account *quality at the manufacturing phase,* and *reliability during system operation,* simultaneously. By rewarding high system reliability, and penalizing the quality loss due to variation, the proposed model can determine the optimal specification limits for inspection, and an optimal replacement interval. Although the idea of integrating quality and reliability into system design is not a new concept, there is no existing approach for determining specifications on degradation characteristics to optimize quality and reliability simultaneously. This integrated strategy can reduce warranty cost, and repair cost, while increasing customer satisfaction in the long run.

Only limited mathematical tools for improving MEMS reliability, quality, and productivity are currently available. This paper proposes a model that can be used as a tool for decision-makers in MEMS manufacturing to economically optimize operational decisions on reliability, quality, and productivity, which are critical factors during the fabrication of the microengine. While illustrated using Sandia-developed microengines as examples [24], [25], the proposed model can be applied to a broad range of MEMS devices that experience wear degradation between rubbing surfaces.

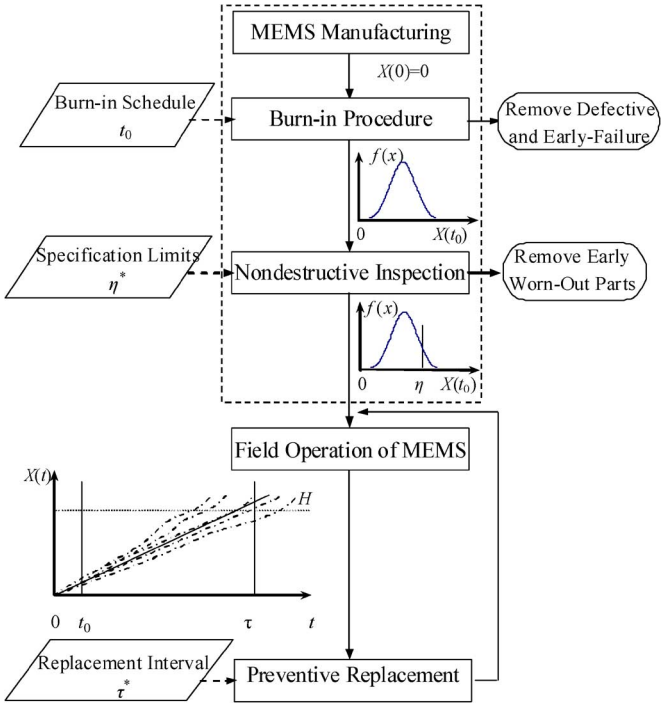


Fig. 2. Burn-in, inspection, and preventive replacement procedures for MEMS.

II. Model Formulation

Consider a MEMS system containing one microengine that is subject to wear. Furthermore, failure of the microengine causes failure of the system. The failure of the microengine occurs when the wear volume of material reaches a critical threshold, H [24]. This type of failure is referred to as "soft" failures, as opposed to "hard" failures when systems or components stop functioning abruptly. The critical threshold on a wear volume is assumed to be a constant in this study, although it may vary from unit-to-unit. The wear volume of material can be estimated by measuring the volume of wear debris, or the missing volume in the worn device. For example, a Focused Ion Beam system is effective to evaluate the amount of wear debris by producing cross sections of the precise area of interest in MEMS structures [25].

To simultaneously improve quality and reliability over the lifetime of MEMS systems, a systematic inspection and preventive replacement procedure has been developed, as depicted in Fig. 2. The initial wear volume of material after the completion of manufacturing is assumed to be zero, i.e., $X(0) = 0$. A burn-in procedure following MEMS manufacturing is used to detect, and remove defective, and early-failed parts. Burn-in is an important process to achieve reliable components and systems, but it also exposes all units to stresses. For the burned-in units, the nondestructive inspection is implemented to screen-out the fraction of units whose wear volumes exceed a certain specification limit. The screened units, with high quality level, are then released for field operation until reaching the periodic replacement time, where the cost of an impending failure makes it economical to replace it with a new one. The preventive replacement procedure is used to prevent failure due to the wear-out of typical operating units.

A. Wear Degradation Model

Let $X(t;\boldsymbol{\beta})$ denote the actual degradation path of a degrading characteristic over time t, where $\boldsymbol{\beta}$ is a vector of model coefficients. The choice of a degradation model requires not only specification of the form of the $X(t;\boldsymbol{\beta})$ function, but also specification of fixed and random parameters in $\boldsymbol{\beta}$ [7]. Typically, degradation paths are described by a model with up to four parameters. Some of the parameters in $\boldsymbol{\beta}$ are random from unit-to-unit, and one or more parameters could be modeled as constant across all units [19].

For the wear degradation of microengines, the degradation model is derived based on physical theory to quantify the functional relationship between the wear volume, $X(t;\boldsymbol{\beta})$, and the number of revolutions to failure, t [26]. Given the radius of the pin joint, r (shown in Fig. 1), the coefficient related to wear and hardness of the material, c, and the force between the contacting surfaces, F, the linear degradation path, $X(t;r,c,F)$, is shown in Fig. 3, and can be expressed as

$$X(t;r,c,F) = 2\pi rcFt. \tag{1}$$

c is a parameter that is directly proportional to the wear coefficient, and inversely proportional to the hardness of material. The radius of pin joint, r, is random from unit-to-unit with mean μ_r, and standard deviation σ_r. For a sinusoidal drive signal, the force applied between rubbing surfaces, F, varies with drive frequency as the critical frequency for resonance is approached. At a given drive frequency, the force applied between rubbing surfaces is random among units with nominal value μ_F, and standard deviation σ_F.

The wear volume at any time t, $X(t)$, is random from unit-to-unit, and can be reasonably assumed to follow a s-normal distribution. There are many combinations of distributions for r, and

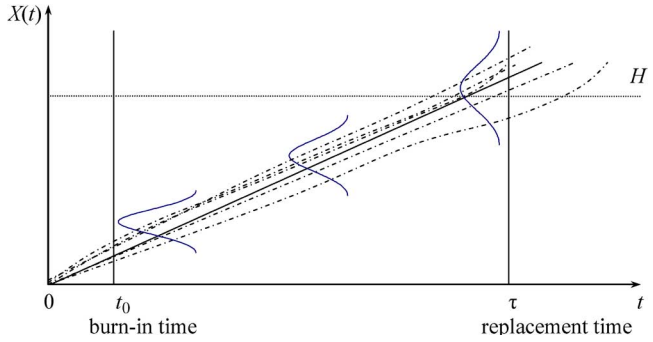


Fig. 3. Linear wear degradation path.

F that will result in a s-normally distributed $X(t)$. If a s-normal distribution is not appropriate, then transformations can be performed to obtain a s-normally distributed random variable. Assuming s-independence between r and F, it is demonstrated that

$$\mu_t = 2\pi c \mu_r \mu_F t, \quad \text{and} \tag{2}$$

$$\sigma_t = 2\pi c t \sqrt{\sigma_r^2 \sigma_F^2 + \sigma_r^2 \mu_F^2 + \mu_r^2 \sigma_F^2}, \tag{3}$$

which indicates that the mean μ_t changes linearly, and the standard deviation σ_t increases linearly, over time. At the completion of MEMS manufacturing, but prior to burn-in, $X(0) = 0$.

For different degradation characteristics in MEMS or other applications, the functional form of $X(t;\boldsymbol{\beta})$, and the approximate transformation may be suggested by physical or chemical theory, past experience, or the available data. When the degradation path is not linear, the scales of $X(t)$, and t can be chosen to simplify the form of the degradation model. For many problems, the Box-Cox family of transformations will be useful, especially the log transformation of degradation and/or time [19].

B. Burn-In Procedure

It is observed through experiments that the occurrence of microengine failures consistently decreases at the early stage of testing, which indicates infant mortality caused by the early failures of defective parts [24]. This implies that a burn-in procedure should be applied following the manufacturing process to effectively remove weak devices from the population. The burn-in process is an extension of manufacturing processes where manufactured units are operated for a short period of time to screen-out defective parts. The burn-in time, t_0, must be determined to prevent the early failures. Selection of the burn-in time is made based on a combination of test data, industry standards, and time restrictions. In this paper, the burn-in time is determined prior to the optimization of the quality tolerance level, and replacement time.

The need for shorter production cycles drives MEMS manufacturers to reduce the burn-in time. However, if the burn-in is incomplete, the microengine may experience unacceptable early failures. An effective burn-in schedule should be determined that it is long enough to induce the defective units to fail, but not too long to impinge on the required lifetime of engines. Further study on burn-in procedure is necessary to incorporate the cost of burn-in procedure into the model, which includes the operating cost of the burn-in equipment, the failure cost during the burn-in process, and marketing losses caused by increased production lead time.

C. Nondestructive Evaluation, and Specification Limits

As wear is the most critical failure mode for the microengine, the wear volume should be carefully evaluated at the end of the burn-in procedure. The units whose wear volumes are beyond a certain specification limit are prone to fail early, and they should be screened to ensure that the wear volume does not unsatisfactorily reduce the lifetime of microengines. Nondestructive evaluation (NDE) systems can be implemented to provide 100% inspection capabilities, such as Focused Ion Beam systems.

At the end of the burn-in procedure, t_0, the wear volume is given as $X(t_0) = 2\pi rcFt_0$, which follows a s-normal distribution, $X(t_0) \sim N(\mu_0, \sigma_0^2)$, where $\sigma_0^2 = (2\pi c)^2(\sigma_r^2\sigma_F^2 + \sigma_r^2\mu_F^2 + \mu_r^2\sigma_F^2)t_0^2$, and $\mu_0 = 2\pi c\mu_r\mu_F t_0$, respectively. During the NDE, an upper specification limit (USL) should be applied based on the wear volume, to screen the units that have a large amount of wear after the burn-in process. The selection of the USL is a crucial decision that is usually determined by optimizing the quality of a system [5], [6]. However, this may be inefficient for the manufacturing of new technologies such as MEMS because of the degradation process, and reliability concerns. Therefore, quality and reliability should be integrated in the optimization of specification limits.

During the NDE, three quality-related costs are considered: quality losses due to the deviation from the ideal value, scrap or rework cost, and inspection cost [5]. The quality loss of each unit can be measured by a quality loss function, which can be chosen based on the type of the quality characteristic: the smaller the better (S-type), the larger the better (L-type), or the target the best (T-type). The wear volume is clearly an S-type quality characteristic, and its quality loss function is expressed as

$$L\left(X(t_0)\right) = kX(t_0)^2, \tag{4}$$

where k is the coefficient that transforms deviations into economic values. The quality loss can be estimated using the expected value of $L(X(t_0))$. Based on the derivation of the expected quality loss for the T-type characteristic provided in Feng & Kapur [6], the expected quality loss for the S-type characteristic, $C_Q(\eta)$, is proven to be

$$\begin{aligned} C_Q(\eta) &= \int_0^{USL} L(x;t_0) f_{X(t_0)}(x;t_0)dx = \int_0^{\eta} kx^2 f_{X(t_0)}(x;t_0)dx \\ &= -k\sigma_0[\mu_0+\eta]\phi\left(\frac{\eta-\mu_0}{\sigma_0}\right) + \sigma_0\mu_0 k\phi\left(\frac{-\mu_0}{\sigma_0}\right) \\ &\quad - k\left[\sigma_0^2+\mu_0^2\right]\Phi\left(\frac{-\mu_0}{\sigma_0}\right) + k\left[\sigma_0^2+\mu_0^2\right]\Phi\left(\frac{\eta-\mu_0}{\sigma_0}\right), \end{aligned} \tag{5}$$

where $f_{X(t_0)}(x;t_0)$ is the pdf of the wear volume at the end of the burn-in process.

If an observed measurement is outside the USL, the unit will be reworked or scrapped. Let $q(\eta)$ be the fraction of

conforming units, which corresponds to the area under the pdf curve bounded by the USL.

$$q(\eta) = \int_0^{\eta} f_{X(t_0)}(x;t_0)dx = \Phi\left(\frac{\eta-\mu_0}{\sigma_0}\right) - \Phi\left(-\frac{\mu_0}{\sigma_0}\right). \quad (6)$$

If the scrap/reworked cost per unit is denoted as s, then the scrapped portion of $(1-q(\eta))$ results in an expected scrap cost of $(1-q(\eta))s$. Thus, the expected scrap cost is

$$C_S(\eta) = (1-q)s = \left(1 - \Phi\left(\frac{\eta-\mu_0}{\sigma_0}\right) + \Phi\left(-\frac{\mu_0}{\sigma_0}\right)\right)s. \quad (7)$$

The inspection cost per unit is denoted as C_I, which is a constant independent of η. Therefore, the total expected quality cost per unit incurred at the manufacturing is expressed as

$$QC(\eta) = C_Q(\eta) + C_S(\eta) + C_I. \quad (8)$$

D. Preventive Replacement, and Cost of Failure

A preventive periodic-replacement policy is used to prevent failure due to the wear-out of typical operating units. As the system ages, it is more economical to replace an aged system because the cost of a planned replacement is less than the associated cost of unscheduled maintenance. The microengine fails when the wear volume of material reaches a critical threshold, H. Therefore, the reliability of a microengine at any time t (or number of cycles) can be assessed by the probability that the wear volume is less than the failure threshold, i.e.,

$$R(t) = P(X(t) < H) = \int_0^{H} f_{X(t)}(x;t)dx = \Phi\left(\frac{H-\mu_t}{\sigma_t}\right). \quad (9)$$

This equation is valid when the wear volume follows a s-normal distribution, i.e., $X(t) \sim N(\mu_t, \sigma_t^2)$. If another distribution is more suitable, an analogous relationship can be derived. The reliability at any time $t(t_0 < t < \tau)$, is measured as a conditional reliability given the probability that the wear volume during the burn-in process is less than H, or

$$R(t|t_0) = \frac{P(X(t) < H)}{P(X(t_0) < H)}, \quad t_0 < t < \tau. \quad (10)$$

To be consistent with the monetary measure of quality costs measure, the system reliability can be evaluated considering the cost-of-failure approach [28]. The cost of failure per unit is assumed to be a constant, f_c, which is s-independent of the time to failure, and can be estimated by a one-year warranty cost, or a one-time repair cost. The system reliability at the time of replacement is then evaluated by the expected failure cost:

$$FC(\tau) = f_c\,(1 - R(\tau|t_0)). \quad (11)$$

If the system fails prior to τ, then it must be replaced by an operational replacement, and the cost is $f_c + RC$, where RC is the replacement cost. Alternatively, if it has not failed by τ, it should be replaced based on economic considerations, and the cost is just RC. Thus, the expected total failure plus replacement cost at τ is $FC(\tau) + RC$.

E. Simultaneous Quality and Reliability Optimization Model

By simultaneously rewarding high reliability during system operation, and penalizing quality loss during manufacturing, a comprehensive model is proposed to determine the specification limit for inspection, η, and the replacement interval, τ. The expected total system cost includes the expected quality cost, failure cost, and replacement cost, which should be minimized over the expected usage time of a microengine. The expected usage time, $E[U|t_0,\tau]$, is demonstrated to be (see the Appendix)

$$E[U|t_0,\tau] = \int_0^{\tau-t_0} R(t+t_0|t_0)dt = \frac{1}{R(t_0)}\left(\tau R(\tau) - t_0 R(t_0) + \int_{t_0}^{\tau} t f_T(t)dt\right), \quad (12)$$

where $f_T(t)$ is the pdf of the failure time with the form

$$f_T(t) = \frac{H}{\sqrt{2\pi}bt^2} e^{-\frac{(H-at)^2}{2b^2t^2}}, \quad (13)$$

with $a = 2\pi c\mu_r\mu_F$ and $b = \sqrt{(2\pi c)^2(\sigma_r^2\sigma_F^2 + \sigma_r^2\mu_F^2 + \mu_r^2\sigma_F^2)}$. This is the pdf for a two-parameter Bernstein distribution [1].

In this way, the expected total system cost per unit expected usage time is given as

$$TC(\eta,\tau) = \frac{QC(\eta) + FC(\tau) + RC}{E[U|t_0,\tau]}. \quad (14)$$

In practice, the upper bound of the replacement interval is usually specified, and is denoted as B_τ. Thus, the constrained optimization model that minimizes the expected total system cost rate due to quality loss, and unreliability during the system life cycle, can be expressed as

$$(\eta^*,\tau^*) = \arg\min\left\{TC(\eta,\tau) = \frac{QC(\eta) + FC(\tau) + RC}{E[U|t_0,\tau]}\right\} \quad \text{subject to} \quad \eta > \mu_0, \quad t_0 \le \tau \le B_\tau. \quad (15)$$

We implemented a sequential quadratic programming (SQP) method to solve the constrained nonlinear problem, because it outperforms many other methods in terms of efficiency, accuracy, and percentage of successful solutions [20].

III. Numerical Examples

Consider a MEMS system with a microengine subject to wear degradation. As given in Tanner *et al.* [24], the coefficient c in (1) is 3×10^{-4} $\mu\mathrm{m}^2/\mathrm{N}$, the mean value of the radius r of the pin joint is 1.5 μm, and the nominal value of the force applied between rubbing surfaces is 3×10^{-6} N. The standard deviations of the radius, and the applied force are assumed to be 5% of their respective mean values. The burn-in period, t_0, is assumed to be 1,000 revolutions. Using (2) and (3), the mean, and standard deviation of the wear volume at the end of the burn-in

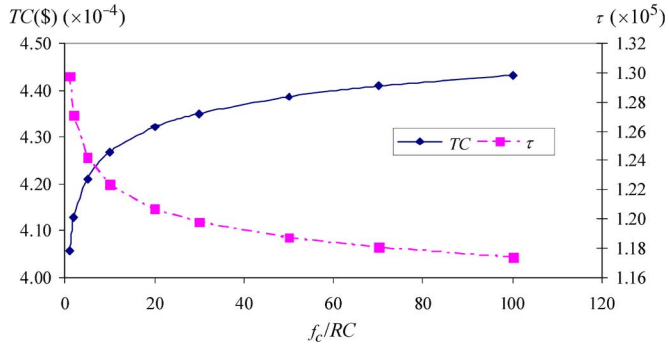


Fig. 4. Sensitivity analysis of TC, and τ, on f_c/RC.

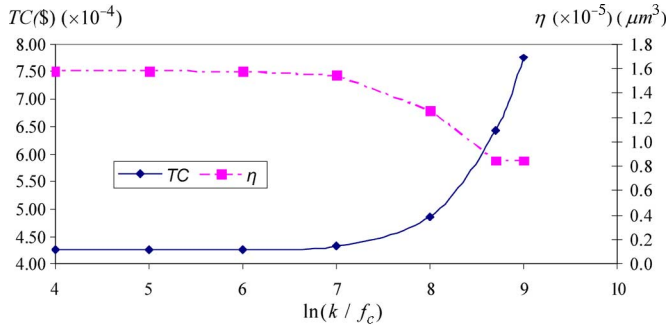


Fig. 5. Sensitivity analysis of TC, and η, on $\ln(k/f_c)$.

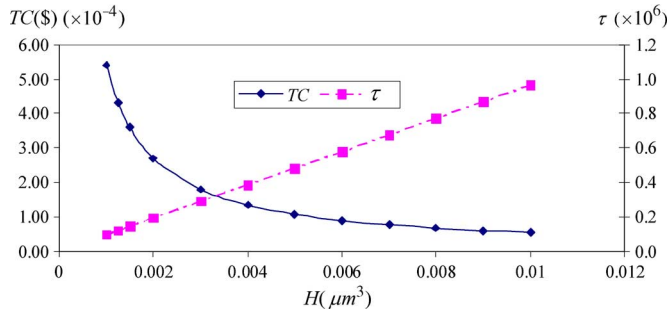


Fig. 6. Sensitivity analysis of TC, and τ, on H.

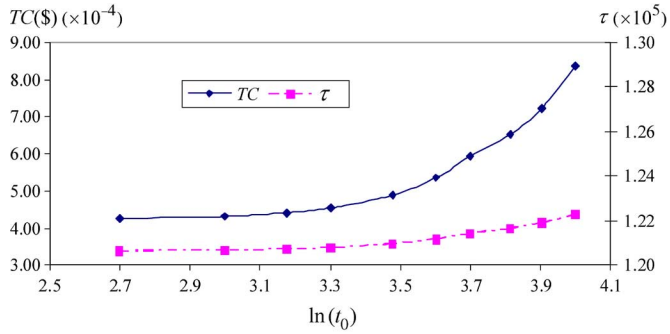


Fig. 7. Sensitivity analysis of TC, and τ, on $\ln(t_0)$.

procedure are 8.4823×10^{-6} μm^3, and 6.0016×10^{-7} μm^3, respectively. The microengine experiences "soft" failures when the wear volume of material reaches a critical threshold, H, which is 0.00125 μm^3. The following cost parameters are used to illustrate this example. The coefficient, k, in the quality loss function is set to be 10^{10}. The cost to inspect the microengine at the manufacturing phase is assumed to be \$0.1 per unit, and the scrap/rework cost of a nonconforming unit is \$20. The replacement cost of the microengine is \$50, and the cost of failure is assumed to be \$1,000. The microengine has to be replaced before 10^5 revolutions, which is the upper bound of the replacement interval.

Using SQP methods provided by MATLAB, the optimal solution is obtained, which indicates that the upper specification limit should be set at 1.5461×10^{-5} μm^3, and the microengine should be replaced every 1.2068×10^5 revolutions. The resultant minimum total cost per expected revolution, TC, is about $\$4.32 \times 10^{-4}$/revolution.

A. *Sensitivity Analysis*

Sensitivity analysis was also performed to observe the effects of model parameters on optimal solutions. The parameters that we are interested in include the ratio between the cost of failure per unit and the replacement cost, f_c/RC; the ratio between the coefficient in the quality loss function and the cost of failure per unit, k/f_c; the critical threshold value, H; and the burn-in time, t_0. The results are shown in Figs. 4 to 7, respectively. It can be observed how the optimal solution changes as each parameter changes.

The ratio f_c/RC indicates the relative magnitude of the failure cost to the replacement cost. When f_c/RC increases from 1 to 100 (keeping RC as a constant) as shown in Fig. 4, the expected total system cost rate increases from $\$4.06 \times 10^{-4}$ to $\$4.43 \times 10^{-4}$, and the optimal replacement interval decreases from 1.2978×10^{-5} to 1.1737×10^5 revolutions. It suggests that microengines should be replaced more frequently as the failure cost increases, while the replacement cost keeps the same.

The ratio between k and f_c represents the relative magnitude between quality loss and failure cost. As shown in Fig. 5, when the ratio between k and f_c increases from 10^4 to 10^9 (keeping f_c as a constant), the expected total system cost rate increases from $\$4.26 \times 10^{-4}$ to $\$7.76 \times 10^{-4}$, and the upper specification limit of wear volume reduces from 1.5820×10^{-5} μm^3 to 8.4823×10^{-6} μm^3. The result indicates that, as k increases, a larger fraction needs to be scrapped or reworked to lower down the cost due to quality loss.

When the critical threshold value, H, increases from 0.001 μm^3 to 0.01 μm^3 as shown in Fig. 6, the expected total system cost rate decreases from $\$5.41 \times 10^{-4}$ to $\$5.36 \times 10^{-5}$, and the replacement interval increases linearly from 9.6547×10^4 to 9.6530×10^5 revolutions. This result suggests that the threshold value has a significant effect on the determination of the replacement interval.

As presented in Fig. 7, when the burn-in time, t_0, increases from 500 to 10,000 revolutions ($\ln(t_0)$ increases from 2.7 to 4), then the replacement interval increases slightly from 1.2065×10^5 to 1.2230×10^5 revolutions, and the total cost per unit increases from $\$4.26 \times 10^{-4}$ to $\$8.38 \times 10^{-4}$. It implies that a shorter burn-in period should be applied to minimize the total system cost, while the burn-in cost is not incorporated into the system cost. It suggests a potential research direction to simultaneously determine the burn-in time while considering the associated cost.

IV. Conclusions

This study proposes a mathematical model to jointly determine inspection and preventive replacement policies for the surface-micromachined microengines subject to wear degradation, which is a major failure mechanism in MEMS devices. For the microengine example, the optimal specification limit for the inspection & the replacement interval are determined by optimizing MEMS quality and reliability simultaneously.

While illustrated using one specific microengine for die-level reliability, the proposed model can be extended to a broader range of MEMS devices that experience wear degradation between rubbing surfaces. For example, a MEMS system with several homogenous or heterogeneous degradation components presents more challenging issues on modeling the interactions between components and system reliability. Furthermore, for the reliability of a final MEMS product, all aspects of fabrication, packaging, system integration, and manufacturing must be considered.

Appendix

Deriving Expected Usage Time, and the pdf of Failure Time

The expected usage time, $E[U|t_0,\tau]$, is

$$\begin{aligned} E[U|t_0,\tau] &= \int_0^{\tau-t_0} R(t+t_0|t_0)dt \\ &= \frac{1}{R(t_0)} \int_0^{\tau-t_0} R(t+t_0)dt \\ &= \frac{1}{R(t_0)} \int_{t_0}^{\tau} R(t')dt' \\ &= \frac{1}{R(t_0)} \left(\tau R(\tau) - t_0 R(t_0) + \int_{t_0}^{\tau} t f_T(t)dt \right). \end{aligned}$$

Using (9), the pdf of the failure time, $f_T(t)$, is derived as

$$\begin{aligned} f_T(t) &= -\frac{dR(t)}{dt} \\ &= -\phi\left(\frac{H-\mu(t)}{\sigma(t)}\right)\left(\frac{1}{\sigma(t)^2}\left(bt(-a)-(H-at)(b)\right)\right) \\ &= \frac{Hb}{b^2t^2}\phi\left(\frac{H-\mu(t)}{\sigma(t)}\right) \\ &= \frac{H}{t\sigma(t)}\phi\left(\frac{H-\mu(t)}{\sigma(t)}\right) \\ &= \frac{H}{\sqrt{2\pi}bt^2}e^{-\frac{(H-at)^2}{2b^2t^2}} \end{aligned}$$

where $a = 2\pi c\mu_r\mu_F$ and $b = \sqrt{(2\pi c)^2(\sigma_r^2\sigma_F^2+\sigma_r^2\mu_F^2+\mu_r^2\sigma_F^2)}$.

Acknowledgment

The authors would like to thank Dr. Danelle Tanner from Sandia National Laboratories for her review of the original manuscript and her insightful comments.

References

[1] M. Ahmad and A. K. Sheikh, "Bernstein reliability model: Derivation and estimation of parameters," *Reliability Engineering*, vol. 8, pp. 131–148, 1984.
[2] S. J. Bae and P. H. Kvam, "A change-point analysis for modeling incomplete burn-in for light displays," *IIE Trans.*, vol. 38, no. 6, pp. 489–498, 2006.
[3] M. Boulanger and L. A. Escobar, "Experiment design for a class of accelerated degradation tests," *Technometrics*, vol. 36, no. 3, pp. 260–272, 1994.
[4] A. Drapella and S. Kosznik, "Combining preventive replacement and burn-in procedures," *Quality and Reliability Engineering International*, vol. 18, no. 5, pp. 423–427, 2002.
[5] Q. Feng, "Integrated Statistical and Optimization Strategies for the Improvement of Six Sigma Methodology," PhD Dissertation, University of Washington, Seattle, WA, 2005.
[6] Q. Feng and K. C. Kapur, "Economic development of specifications for 100% inspection based on asymmetric quality loss functions," *IIE Trans.*, vol. 38, no. 8, pp. 659–669, 2006.
[7] Q. Feng and D. W. Coit, "Simultaneous quality and reliability optimization for systems composed of degrading components," in *Proceedings of Industrial Engineering Research Conference*, Nashville, TN, May 19–23, 2007.
[8] A. Grall, L. Dieulle, C. Berenguer, and M. Roussignol, "Continuous-time predictive-maintenance scheduling for a deteriorating system," *IEEE Trans. Reliability*, vol. 51, no. 2, pp. 141–150, 2002.
[9] R. Jiang and A. K. S. Jardine, "An optimal burn-in preventive-replacement model associated with a mixture distribution," *Quality and Reliability Engineering International*, vol. 23, pp. 83–93, 2007.
[10] V. R. Joseph and I. T. Yu, "Reliability improvement experiments with degradation data," *IEEE Trans. Reliability*, vol. 55, no. 1, pp. 149–157, 2006.
[11] J. P. Kharoufeh, "Explicit results for wear processes in a Markovian environment," *Operations Research Letters*, vol. 31, pp. 237–244, 2003.
[12] J. P. Kharoufeh and S. M. Cox, "Stochastic models for degradation-based reliability," *IIE Trans.*, vol. 37, pp. 533–542, 2005.
[13] W. Kuehnel and S. Sherman, "A surface micromachined silicon accelerometer with on-chip detection circuitry," *Sensors and Actuators*, vol. 45, no. 1, pp. 7–16, 1994.
[14] W. Kuo, "Challenges related to reliability in Nano electronics," *IEEE Trans. Reliability*, vol. 55, no. 4, pp. 569–570, 2006.
[15] H. Liao, E. A. Elsayed, and L. Y. Chan, "Maintenance of continuously monitored degrading systems," *European Journal of Operational Research*, vol. 175, pp. 821–835, 2006.
[16] C. J. Lu and W. Q. Meeker, "Using degradation measures to estimate a time-to-failure distribution," *Technometrics*, vol. 35, no. 2, pp. 161–174, 1993.
[17] H. Lu, W. J. Kolarik, and S. S. Lu, "Real-time performance reliability prediction," *IEEE Trans. Reliability*, vol. 50, no. 4, pp. 353–357, 2001.
[18] S. Lu, Y. C. Tu, and H. Lu, "Predictive condition-based maintenance for continuously deteriorating systems," *Quality and Reliability Engineering International*, vol. 23, pp. 71–81, 2007.
[19] W. Q. Meeker, L. A. Escobar, and C. J. Lu, "Accelerated degradation tests: Modeling and analysis," *Technometrics*, vol. 40, no. 2, pp. 89–99, 1998.
[20] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York: Springer-Verlag, 2006.
[21] K. A. Peterson, P. Tangyunyong, and A. A. Pimentel, "Failure analysis of surface-micromachined microengines," in *Proceedings of SPIE, the Materials and Device Characterization in Micromachining Symposium*, C. R. Friedrich and Y. Vladimirsky, Eds., Santa Clara, CA, Sep. 21–22, 1998, vol. 3512, pp. 190–200.
[22] A. D. Romig and P. J. McWhorter, "Intelligent micromachines: Opportunities and challenges of the next Si revolution (invited paper)," *Semicon*, 2001a, Europe.
[23] A. D. Romig and P. J. McWhorter, "Opportunities and challenges in MEMS commercialization," *Vacuum Coat Technology Magazine*, 2001b.

[24] D. M. Tanner, W. M. Miller, K. A. Peterson, M. T. Dugger, W. P. Eaton, L. W. Irwin, D. C. Senft, N. F. Smith, P. Tangyunyong, and S. L. Miller, "Frequency dependence of the lifetime of a surface micromachined microengine driving a load," *Microelectronics Reliability*, vol. 39, pp. 401–414, 1999a.

[25] D. M. Tanner, J. A. Walraven, L. W. Irwin, M. T. Dugger, N. F. Smith, W. M. Miller, and S. L. Miller, "The effect of humidity on the reliability of a surface micromachined microengine," in *Proceedings of IEEE International Reliability Physics Symposium*, 1999b, pp. 189–197.

[26] D. M. Tanner and M. T. Dugger, "Wear mechanisms in a reliability methodology," in *SPIE's Proceedings*, San Jose, CA, 2003, vol. 4980, Reliability, Testing, Characterization of MEMS/MOEMS, pp. 22–40.

[27] S. T. Tseng, M. Hamada, and C. H. Chiao, "Using degradation data from a factorial experiment to improve fluorescent lamp reliability," *Journal of Quality Technology*, vol. 27, no. 4, pp. 363–369, 1995.

[28] M. T. Todinov, "Reliability analysis and setting reliability requirements based on the cost of failure," *International Journal of Reliability, Quality and Safety Engineering*, vol. 11, no. 3, pp. 273–299, 2004.

[29] N. Unal and R. Weschsung, "Inkjet printheads: An example of MST market reality," *Micromachine Devices*, vol. 3, no. 1, pp. 1–6, 1998.

[30] J. A. Walraven, T. J. Headley, A. N. Campbell, and D. M. Tanner, "Failure analysis of worn surface micromachined microengines," in *SPIE Proceedings on Micromachining and Microfabrication*, Santa Clara, CA, September 20–22, 1999, pp. 30–40.

[31] H. Yu and C. H. Chiao, "An optimal designed degradation experiment for reliability improvement," *IEEE Trans. Reliability*, vol. 51, no. 4, pp. 427–433, 2002.

Hao Peng is a Ph.D. Student in the Department of Industrial Engineering at the University of Houston. She received a BS degree in industrial engineering from Tsinghua University, Beijing, China in 2006. Her research interests include reliability & maintenance engineering, and optimization, especially degradation-based modeling and analysis. She is a member of INFORMS, and ASQ.

Qianmei Feng is an Assistant Professor in the Department of Industrial Engineering at the University of Houston. She received a Ph.D. degree in industrial engineering from the University of Washington, Seattle, WA in 2005. Her research has been in reliability and quality engineering; and applications in manufacturing, healthcare, and transportation systems. She has published a dozen papers in journals such as *IIE Transactions, Reliability Engineering and System Safety, International Journal of Advanced Manufacturing Technology,* and *Risk Analysis*. She is a member of IIE, INFORMS, ASQ, and Alpha Pi Mu.

David W. Coit (M'03) is an Associate Professor in the Department of Industrial & Systems Engineering at Rutgers University. He received a BS degree in mechanical engineering from Cornell University, an MBA from Rensselaer Polytechnic Institute, and MS & Ph.D. degrees in industrial engineering from the University of Pittsburgh. In 1999, he was awarded a CAREER grant from NSF to study reliability optimization. He also has over ten years of experience working for IIT Research Institute (IITRI), Rome, NY, where he was a reliability analyst, project manager, and an engineering group manager. His current research involves reliability prediction & optimization, risk analysis, and multi-criteria optimization considering uncertainty. He is a member of IIE, and INFORMS.

Reliability Modeling for MEMS Devices Subjected to Multiple Dependent Competing Failure Processes

Hao Peng, Qianmei Feng
Department of Industrial Engineering
University of Houston, Houston, TX 77204, USA

David W. Coit
Department of Industrial and Systems Engineering
Rutgers University, Piscataway, NJ 08854, USA

Abstract

Widespread acceptance of micro-electro-mechanical systems (MEMS), both for large-volume commercialization and for critical applications, depends highly on their reliability. Most previous studies in MEMS reliability focused on identification of failure mechanisms and development of predictive models for such mechanisms. We develop a preventive maintenance model for MEMS devices subjected to multiple dependent competing failure processes including shock loads and a degradation process to minimize the average cost rate. A numerical example is provided to demonstrate applications for the new model.

Keywords

MEMS reliability, degradation, random shocks, preventive maintenance, dependent competing failure processes

1. Introduction

Reliability studies of micro-electro-mechanical systems (MEMS) are becoming increasingly important to achieve widespread acceptance of such technologies, both for large-volume commercialization and for critical applications. MEMS reliability problems are challenging since each device has its own unique failure modes and mechanisms [1]. Previous studies in MEMS reliability focused on identification of failure mechanisms and development of predictive models for such mechanisms [2]. The common MEMS failure mechanisms and causes are identified as wear degradation, stiction, shock loads and fatigue, among others.

For MEMS devices that have moving parts with rubbing and impacting surfaces, such as microengines and Torsional Ratcheting Actuators (TRA), experiments have been performed to investigate failure modes, particularly in extreme environments (e.g., shock, vibration) [3]. For these MEMS devices, the most prominent failure mechanism is discovered to be the wearing process on rubbing surfaces. Based on a wear degradation model, we developed an integrated quality and reliability optimization model to determine the operational parameters in burn-in, quality inspection and preventive maintenance [4]. This research makes a contribution to the enhancement of quality and reliability for MEMS devices from manufacturing and maintenance perspectives.

MEMS can be exposed to shock environments during fabrication, deployment or operation, which may lead to catastrophic failures or cumulative damages. For instance, accidental drops onto hard surfaces might cause significant reliability problems for microsystems used in automotive, industrial or other critical applications [5]. Random shocks can cause or lead to failures of MEMS in different modes, such as fracture, delamination, and stiction. While a few experimental studies of mechanical responses and reliability were conducted for particular shock-loaded microsystems [5] and microengines [6], no existing research has been performed to use stochastic shock models for the design of dynamically reliable MEMS.

In the research described in this article, probability models are developed based on the combination of random-shock and degradation modeling. The objective of this work is to perform reliability analysis and modeling for MEMS devices subjected to multiple dependent failure processes including both the wear degradation process and random shocks from the operational environment, and to determine a cost-effective preventive maintenance policy

to minimize the average total maintenance cost rate. The models developed in this research can be applied directly or customized for most current and evolving MEMS designs with multiple failure processes.

Degradation modeling has been successfully applied to many applications. As an effective alternative to compensate for insufficient failure data, it can provide a greater understanding of physics/chemical failure modes and mechanisms. It also offers an indirect way to estimate failure-time distributions and to predict reliability. The wear on rubbing surfaces is one example of MEMS degradation. Previous studies on degradation focused on developing degradation path models and estimating time-to-failure distributions [7,8], and investigating the degradation process when exposed to a random environment [9,10]. Reliability modeling of both degradation process and random shocks has been investigated by several researchers. Cumulative damage caused by random shocks, as well as a degradation process, compose two competing failure processes in some proposed models [11-13]. Catastrophic failures caused by random shocks (such as interface fracture) and aging degradation failures were simultaneously analyzed using probability models [14, 15].

According to Sandia's reliability tests on micro-actuator systems (e.g., microengines, TRA) [6], both cumulative damage caused by wear degradation as well as random shocks, and catastrophic failures caused by random shocks, have a significant impact on the reliability of MEMS. These two competing failure processes are dependent in a way that the same random shock process impacts both cumulative wear damage and catastrophic failures. This case has not been previously studied in existing research, but will be specifically addressed in our reliability analysis for enhancing MEMS reliability. Furthermore, a preventive maintenance policy is adopted to establish an optimal maintenance interval for MEMS, since a cost-effective preventive maintenance policy can significantly extend a system's life and reduce the number of failures compared to a corresponding corrective maintenance policy [16].

2. Modeling of Multiple Dependent Failure Processes

Consider a micro-actuator system that may fail due to two competing yet dependent failure modes: (1) soft failures caused by wear degradation and cumulative wear damages from a random shock process, and (2) catastrophic failures caused by extreme shocks from the same random shock process. The system fails when either of the two competing failure processes reaches its threshold value.

Soft failure occurs when the overall wear volume is beyond a threshold value H. The wear volume is caused by aging degradation and cumulative damages (in the form of debris) due to random shocks, according to a *cumulative shock* model [11]. The wear volume due to degradation over time t often follows a linear degradation path [4]. In addition, the wear volume increases instantaneously in the form of debris when a shock arrives. The wear damage caused by the random shock and the linear degradation path are assumed to be independent.

Hard/catastrophic failure occurs when the shock load/stress exceeds the maximal fracture strength D, according to an *extreme shock* model [6]. Random shocks arrive according to a Poisson process. The sizes of shock loads/stresses are independently and identically distributed (i.i.d.) normal random variables. The magnitudes of shock damage impacting wear volume are also i.i.d. normal variables.

In our model, no continuous monitoring is performed on the system. The system is inspected at periodic intervals, τ, and the inspection determines whether the system is operational or not. If the system is operational, the inspection also detects the wear volume and compares it with the soft failure threshold H. Inspections are assumed to be instantaneous, perfect, and non-destructive. If the system is detected to be failed, it is replaced instantly with a new one. If the system is operating, it is kept undisturbed until next inspection period.

The evolution of the system condition over time is illustrated in Figure 1 including the wear degradation process (top figure) and the random shock load process (bottom figure). The details are further discussed in the following sections.

2.1 Extreme Shock Model for Catastrophic Failures

Assume shocks arrive according to a Poisson process $\{N(t), t\geq 0\}$ with rate λ. As shown in Figure 1 (bottom), shock loads or applied stresses arriving at $t_1, t_2, \ldots, t_n$ are denoted as $\{W_1, W_2, \ldots, W_n\}$, where n is the number of shocks occurring during the degradation process. $\{W_1, W_2, \ldots, W_n\}$ are assumed to be i.i.d. random variables distributed as a normal distribution, $W_i \sim N(\mu_w, \sigma_w^2)$. The device experiences a hard failure when the applied stress exceeds the fracture strength D. For each shock, the probability that the device survives the applied stress is

$$P_L = P(W_i < D) = \Phi\left(\frac{D-\mu_W}{\sigma_W}\right). \tag{1}$$

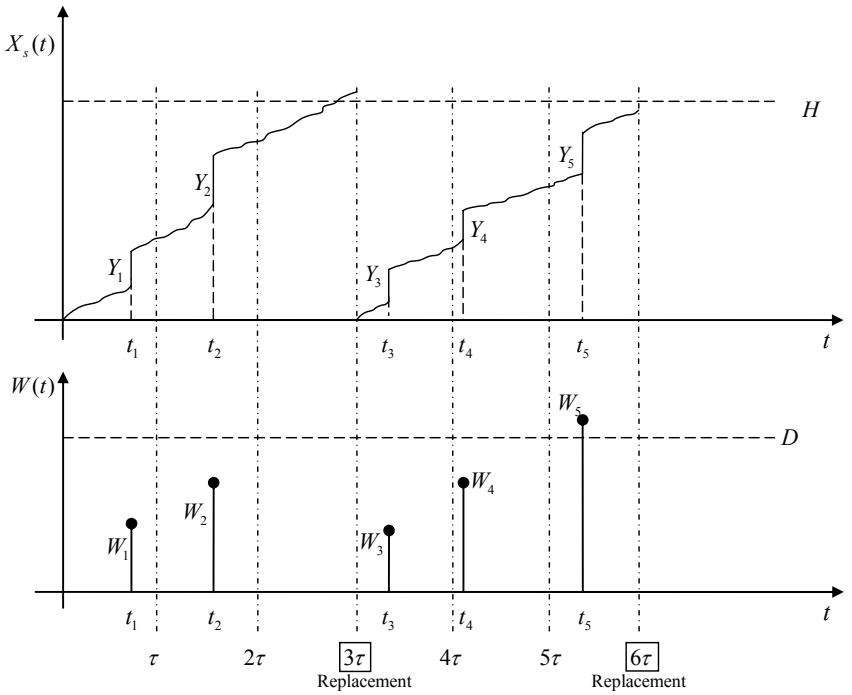


Figure 1: System condition over time

2.2 Cumulative Shock Model for Soft Degradation Failures

The wear degradation paths $X_s(t)$ are illustrated in Figure 1 (top), which results from both aging degradation and instantaneous damages (in the form of debris) due to random shocks. The aging wear process follows a linear degradation path $X(t)=\varphi+\beta t$, where the degradation rate β follows a normal distribution and the initial value φ is a constant [4]. Every random shock during operation causes an instantaneous step increase on the wear volume, measured by the shock damage size. Shock damage sizes are assumed to be random i.i.d variables, denoted as $\{Y_1, Y_2, \ldots, Y_n\}$, where n is the number of shocks appeared during the degradation process. The cumulative damage size due to random shocks until time t, $S(t)$, is given as

$$S(t) = \begin{cases} \sum_{i=1}^{N(t)} Y_i, \text{if } N(t) > 0 \\ 0, \quad \text{if } N(t) = 0. \end{cases} \tag{2}$$

Therefore, the overall degradation path $X_s(t)$ considering both linear degradation and random shock damages is expressed as $X_s(t) = X(t) + S(t)$. The cumulative distribution function (cdf) of $X_s(t)$ can be derived as

$$\begin{aligned} F_X(x,t) &= P(X_s(t) < x) = P(X(t) + S(t) < x) = P(\varphi + \beta t + S(t) < x) \\ &= \sum_{n=0}^{\infty} P\left(\varphi + \beta t + S(t) < x \mid N(t) = n\right) P\left(N(t) = n\right). \end{aligned} \tag{3}$$

If we further assume that shock damage sizes $\{Y_1, Y_2, \ldots, Y_n\}$ are i.i.d. normal random variables, $Y_i \sim N(\mu_y, \sigma_y^2)$, and β follows a normal distribution, $\beta \sim N(\mu_\beta, \sigma_\beta^2)$, then

$$F_X(x,t) = \sum_{n=1}^{\infty} \Phi\left(\frac{x-(\mu_\beta t+\varphi+n\mu_Y)}{\sqrt{\sigma_\beta^2 t^2 + n\sigma_Y^2}}\right)\frac{\exp(-\lambda t)(\lambda t)^n}{n!} + \Phi\left(\frac{x-\mu_\beta t-\varphi}{\sigma_\beta t}\right)\exp(-\lambda t). \tag{4}$$

Soft failure occurs when the degradation variable $X_s(t)$ is larger than the failure threshold H. Thus, the probability that there is a soft failure before t is expressed as $P(X_s(t) > H) = 1-F_x(H,t)$.

2.3 System Reliability Analysis

Consider the two competing failure processes, i.e., catastrophic failures and soft degradation failures, the system reliability function can now be expressed as follows:

$$\begin{aligned} R(t) &= 1 - F_T(t) = \sum_{n=0}^{\infty} P_L{}^n P(\varphi + \beta t + S(t) < H \mid N(t) = n) P(N(t) = n) \\ &= \sum_{n=1}^{\infty} P_L{}^n \Phi\left(\frac{H - (\mu_\beta t + \varphi + n\mu_Y)}{\sqrt{\sigma_\beta^2 t^2 + n\sigma_Y^2}} \right) \frac{\exp(-\lambda t)(\lambda t)^n}{n!} + \Phi\left(\frac{H - \mu_\beta t - \varphi}{\sigma_\beta t} \right) \exp(-\lambda t). \end{aligned} \tag{5}$$

Based on the reliability function, the probability density function (pdf) of the failure time distribution is derived as

$$\begin{aligned} f_T(t) &= \frac{dF_T(t)}{dt} = -\frac{dR(t)}{dt} \\ &= -\sum_{n=1}^{\infty} P_L{}^n \phi\left(\frac{H - (\mu_\beta t + \varphi + n\mu_Y)}{\sqrt{\sigma_\beta^2 t^2 + n\sigma_Y^2}} \right) \left(\frac{-\mu_\beta \left(\sigma_\beta^2 t^2 + n\sigma_Y^2\right) - \sigma_\beta^2 t \left(H - (\mu_\beta t + \varphi + n\mu_Y)\right)}{\left(\sigma_\beta^2 t^2 + n\sigma_Y^2\right)^{\frac{3}{2}}} \right) \frac{\exp(-\lambda t)(\lambda t)^n}{n!} \\ &\quad -\sum_{n=1}^{\infty} P_L{}^n \Phi\left(\frac{H - (\mu_\beta t + \varphi + n\mu_Y)}{\sqrt{\sigma_\beta^2 t^2 + n\sigma_Y^2}} \right) \frac{\lambda \exp(-\lambda t)(\lambda t)^{n-1}(-\lambda t + n)}{n!} - \phi\left(\frac{H - \mu_\beta t - \varphi}{\sigma_\beta t} \right) \left(\frac{-H + \varphi}{\sigma_\beta t^2} \right) \exp(-\lambda t) \\ &\quad + \lambda \Phi\left(\frac{H - \mu_\beta t - \varphi}{\sigma_\beta t} \right) \exp(-\lambda t). \end{aligned} \tag{6}$$

2.4 Cost Rate Evaluation for Preventive Maintenance

For non-repairable systems, such as most micro-actuator systems, a periodic inspection maintenance policy can be implemented to minimize the impact of unscheduled failures and to minimize cost. The system is inspected at periodic intervals. If it is detected to be failed, it is replaced instantly with a new one. If the system is operating, it remains undisturbed until next inspection period. To optimize this preventive maintenance policy, the time interval τ for periodic inspection needs to be determined.

We propose an average long-run maintenance cost rate model to evaluate the cost of the preventive maintenance policy, assuming that the time horizon is infinity. From basic renewal theory, the average long-run total maintenance cost per unit time, $\lim_{t\to\infty}(C(t)/t)$, can be evaluated through the first renewal cycle, i.e., $\lim_{t\to\infty}(C(t)/t) = E[TC]/E[K]$, where $C(t)$ is the total maintenance cost during time period t, K is the first cycle length shown in Figure 1, and TC is the total maintenance cost of the first replacement cycle. The expected total maintenance cost per cycle is given as:

$$E(TC) = C_I E[N_I] + C_F E[\rho] + C_R, \tag{7}$$

where C_I is the cost associated with each inspection, C_F is the penalty cost rate during downtime, and C_R is the replacement cost. The expected number of inspections $E[N_I]$ in the first renewal cycle is expressed as

$$E[N_I] = \sum_{i=1}^{\infty} iP(N_I = i) = \sum_{i=1}^{\infty} iP((i-1)\tau < T \le i\tau) = \sum_{i=1}^{\infty} i\left(F_T(i\tau) - F_T\left((i-1)\tau\right)\right), \tag{8}$$

where τ is the time interval of periodic inspection, T is the system failure time due to the dependent competing failure processes, and the pdf of T is given in Equation (6). ρ in Equation (7) denotes the time from a system failure to the next inspection when the failure is then detected, or the system downtime. The expected time length of system downtime $E[\rho]$ in the first renewal cycle is given as

$$E[\rho] = \sum_{i=1}^{\infty} E[\rho \mid N_I = i] P(N_I = i) = \sum_{i=1}^{\infty} \left(\int_{(i-1)\tau}^{i\tau} (i\tau - t) dF_T(t) \right) \left(F_T(i\tau) - F_T\left((i-1)\tau\right)\right). \tag{9}$$

The expected time length of the renewal cycle $E[K]$ is derived as

$$E[K] = \sum_{i=1}^{\infty} E[K \mid N_I = i] P(N_I = i) = \sum_{i=1}^{\infty} i\tau \left(F_T(i\tau) - F_T\left((i-1)\tau\right)\right). \tag{10}$$

Based on Equations (7)-(10), the average long-run maintenance cost rate as a function of τ is given as

$$CR(\tau)=\lim_{t\to\infty}(C(t)/t)=\frac{E[TC]}{E[K]}=\frac{C_I E[N_I]+C_F E[\rho]+C_R}{E[K]}$$

$$=\frac{C_I\left(\sum_{i=1}^{\infty} i\left(F_T(i\tau)-F_T\left((i-1)\tau\right)\right)\right)+C_F\left(\sum_{i=1}^{\infty}\left(\int_{(i-1)\tau}^{i\tau}(i\tau-t)dF_T(t)\right)\left(F_T(i\tau)-F_T\left((i-1)\tau\right)\right)\right)+C_R}{\sum_{i=1}^{\infty} i\tau\left(F_T(i\tau)-F_T\left((i-1)\tau\right)\right)}. \quad (11)$$

By minimizing $CR(\tau)$ using analytical or numerical methods, we can obtain the optimal time interval of periodic inspection for the preventive maintenance policy.

3. Numerical Example

To illustrate the models proposed in this paper, we provide a case study of a micro-actuator system designed by Sandia National Laboratory [6]. As shown in Figure 2, a microengine consists of orthogonal linear comb drive actuators that are mechanically connected to a rotating gear. The linear displacement of the comb drives is transformed to the gear via a pin joint. The gear rotates about a hub that is anchored to the substrate. The dominant failure mechanism is identified as the visible wear on rubbing surfaces, which often results in seized microengines or broken pin joints [6]. The wear volume is primarily caused by the aging degradation process. In addition, shock tests on microengines reveal that shock loads may cause substantial wear debris, as well as spring fracture [6]. Therefore, microengines experience these two competing failure processes: soft failures due to aging degradation and debris from shock loads, and catastrophic failures due to spring fracture. The model proposed in this paper was used to study the microengine reliability and to determine its preventive maintenance policy that can be used in practice.

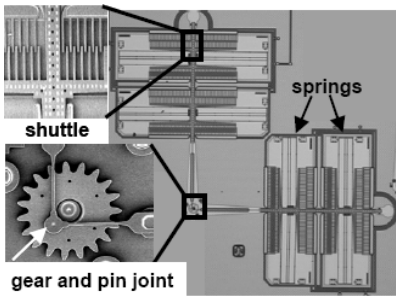


Figure 2: Scanning electron microscopy image of a microengine [6]

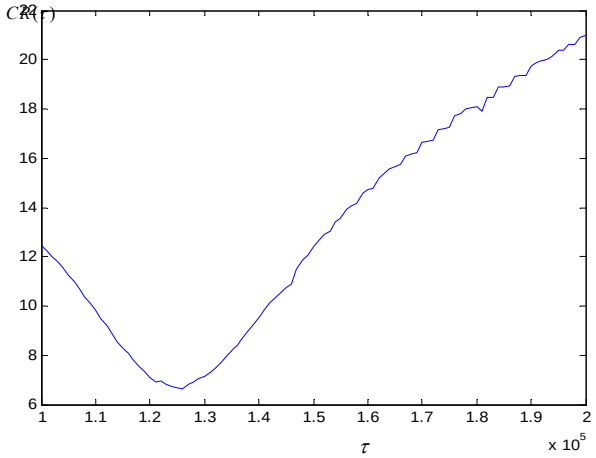


Figure 3: Cost rate versus inspection interval

As studied in [4], the linear degradation path of wear volume is $X(t)=\varphi+\beta t$, in which $\varphi=0$ and the normal random variable β has parameters $\mu_\beta=8.4823\times10^{-9}$ and $\sigma_\beta=6.0016\times10^{-10}$ (t is the number of revolutions in microengines' rotation). The microengine experiences "soft" failures when the wear volume of material reaches a critical threshold, H, which is 0.00125 μm^3 [4, 6]. Random shocks are assumed to follow a Poisson process with rate $\lambda=2.5\times10^{-5}$.

Shock damage sizes $\{Y_1, Y_2, \ldots, Y_n\}$ are i.i.d. normal random variables, with $\mu_y = 1\times10^{-4}$ and $\sigma_y = 2\times10^{-5}$. For polysilicon, the material used for springs, a conservative estimate of the fracture strength is 1.5 GPa [6]. Assume that the parameters for the applied stresses $\{W_1, W_2, \ldots, W_n\}$ (i.i.d. normal random variables) are $\mu_w = 1.2$ GPa and $\sigma_w = 0.2$ GPa, then the probability P_L that the device survives an applied stress for a shock can be calculated by Equation (1) and it equals 93.32%. Set $C_I = 1$, $C_F = 50$, and $C_R = 1$, we obtain the plot of function $CR(\tau)$ as a function of τ in Figure 3. Through numerical calculation, it is easy to find that $C(1.263\times10^5) = 6.6103$ is the minimum average cost rate. In other words, the optimal number of revolutions for periodical inspection is $\tau^* = 1.263\times10^5$. We can also observe from Figure 3 that the optimal value τ^* is uniquely existent.

4. Conclusions

We analyzed reliability and studied the maintenance policy for MEMS devices subjected to multiple dependent failure processes. These failure processes include instantaneous wear damages and catastrophic failures caused by random shocks, as well as aging wear degradation process. A cost-effective preventive maintenance policy is established based on the reliability analysis, and the model is illustrated using a case study on microengines.

References

1. Tanner, D.M., Parson, T.B., Corwin, A.D., Walraven, J.A., Wittwer, J.W., Boyce, B.L., and Winzer, S.R., 2007, "Science-based MEMS reliability methodology," Microelectronic Reliability, 47, 1806-1811.
2. Mason, R., Gintert, L., Rippen, M., Skelton, D., Zunino, J., and Gutmanis, I., 2006, "Guidelines for Reliability Testing of Microelectromechanical Systems in Military Applications," in Reliability, Packaging, Testing, and Characterization of MEMS/MOEMS.
3. Miller, S.L., Rodgers, M.S., Vigne, G.L., Sniegowski, J.J., Clews, P., Tanner, D.M., and Peterson, K.A., 1998, "Failure Modes in surface Micromachined MicroElectroMechanical Actuators," in 36th Annual International Reliability Physics Symposium, Reno, Nevada, 17-25.
4. Peng, H., Feng, Q., and Coit, D.W., 2009, "Simultaneous Quality and Reliability Optimization for Microengines Subject to Degradation," IEEE Transactions on Reliability (In press).
5. Srikar, V.T., and Senturia, S.D., 2002, "The reliability of microelectromechanical systems (MEMS) in shock environments," Journal of Mictoelectromechanical systems, 11, 206-214.
6. Tanner, D.M., Smith, N.F., Irwin L.W., Eaton, W.P., Helgesen, K.S., Clement, J.J., Miller, W.M., Walraven, J.A., Peterson, K.A., Tangyunyong, P., Dugger, M.T., and Miller, S.L., 2000, "MEMS Reliability: Infrastructure, Test Structures, Experiments, and Failure Modes," Sandia National Laboratories, Albuquerque, NM 87185-1081.
7. Lu, C.J., and Meeker, W.Q., 1993, "Using degradation measures to estimate a time-to-failure distribution," Technometrics, 35, 161-174.
8. Bae, S.J., Kuo, W., and Kvam, P.H., 2007, "Degradation models and implied lifetime distributions," Reliability Engineering & System Safety, 92, 601-608.
9. Singpurwalla, N.D., 1995, "Survival in Dynamic Environments," Statistical Science, 10, 86-103.
10. Kharoufeh, J.P., 2003, "Explicit results for wear processes in a Markovian environment," Operations Research Letters, 31, 237-244.
11. Klutke, G.A., and Yang, Y., 2002, "The availability of inspected systems subject to shocks and graceful degradation," IEEE Transactions on Reliability, 51, 371-374.
12. Kharoufeh, J.P., Finkelstein, D.E., and Mixon, D.G., 2006, "Availability of periodically inspected systems with markovian wear and shocks," Journal of Applied Probability, 43, 303-317.
13. Li, W., and Pham, H., 2005, "An inspection-maintenance model for systems with multiple completing processes," IEEE Transactions on Reliability, 54, 318-327.
14. Huang, W., and Askin, R.G., 2003, "Reliability analysis of electronic devices with multiple competing failure modes involving performance aging degradation," Quality and Reliability Engineering International 19, 241-254.
15. Yang, K. and Xue, J., 1996, "Continuous state reliability analysis," in Reliability and Maintainability Symposium, 251-257.
16. Lu, S., Tu, Y.C., and Lu, H., 2007, "Predictive condition-based maintenance for continuously deteriorating systems," Quality and Reliability Engineering International, 23, 71-81.

High-resolution electrohydrodynamic jet printing

JANG-UNG PARK[1], MATT HARDY[1], SEONG JUN KANG[1,2], KIRA BARTON[3], KURT ADAIR[3], DEEP KISHORE MUKHOPADHYAY[3], CHANG YOUNG LEE[4], MICHAEL S. STRANO[4], ANDREW G. ALLEYNE[3], JOHN G. GEORGIADIS[3], PLACID M. FERREIRA[3] AND JOHN A. ROGERS[1,3]*

[1]Department of Materials Science and Engineering, Beckman Institute, and Frederick Seitz Materials Research Laboratory, University of Illinois at Urbana-Champaign, 1304 West Green Street, Urbana, Illinois 61801, USA

[2]Division of Advanced Technology, Korea Research Institute of Standards and Science, 1 Doryong-Dong, Yuseong-Gu, Daejon 305-340, South Korea

[3]Department of Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA

[4]Department of Chemical & Biomolecular Engineering, and Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA

*e-mail: jrogers@uiuc.edu

Published online: 5 August 2007; doi:10.1038/nmat1974

Efforts to adapt and extend graphic arts printing techniques for demanding device applications in electronics, biotechnology and microelectromechanical systems have grown rapidly in recent years. Here, we describe the use of electrohydrodynamically induced fluid flows through fine microcapillary nozzles for jet printing of patterns and functional devices with submicrometre resolution. Key aspects of the physics of this approach, which has some features in common with related but comparatively low-resolution techniques for graphic arts, are revealed through direct high-speed imaging of the droplet formation processes. Printing of complex patterns of inks, ranging from insulating and conducting polymers, to solution suspensions of silicon nanoparticles and rods, to single-walled carbon nanotubes, using integrated computer-controlled printer systems illustrates some of the capabilities. High-resolution printed metal interconnects, electrodes and probing pads for representative circuit patterns and functional transistors with critical dimensions as small as 1 μm demonstrate potential applications in printed electronics.

Printing approaches used in the graphic arts, particularly those based on ink-jet techniques, are of interest for applications in high-resolution manufacturing owing to attractive features that include (1) the possibility for purely additive operation, in which functional inks are deposited only where they are needed, (2) the ability to pattern directly classes of materials such as fragile organics or biological materials that are incompatible with established patterning methods such as photolithography, (3) the flexibility in choice of structure designs, where changes can be made rapidly through software-based printer-control systems, (4) compatibility with large-area substrates and (5) the potential for low-cost operation[1–3]. Conventional devices for ink-jet printing rely on thermal or acoustic formation and ejection of liquid droplets through nozzle apertures[3]. A growing number of reports describe adaptations of these devices with specialized materials in ink formats for applications in electronics[4–9], information display[10–12], drug discovery[13,14], micromechanical devices[15,16] and other areas[3,17]. The functional resolution in these applications, as defined by the narrowest continuous lines or smallest gaps that can be created reliably, is ~20–30 μm (refs 9,18–20). This coarse resolution results from the combined effects of droplet diameters that are usually no smaller than ~10–20 μm (2–10 pl volumes) and placement errors that are typically ±10 μm at standoff distances of ~1 mm (refs 5,21,22). Some methods can avoid these limitations, for certain classes of features. For example, lithographically predefined assist features[5,23,24] or surface functionalization of pre-printed inks[25] in the form of patterns of wettability or surface relief can confine and guide the flow of the droplets as they land on the substrate. In this manner, gaps between printed droplets, for example, can be controlled at the submicrometre level[23–25]. This capability is important for applications in electronics when such gaps define transistor channel lengths. These methods do not, however, offer a general approach to high resolution. In addition, they require separate patterning systems and processing steps to define the assist features.

Electrohydrodynamic jet (e-jet) printing is a technique that uses electric fields, rather than thermal or acoustic energy, to create the fluid flows necessary for delivering inks to a substrate. This approach has been explored for modest-resolution applications (dot diameters ≥20 μm using nozzle diameters ≥50 μm) in the graphic arts[26–29]. To our knowledge, it is unexamined for its potential to provide high-resolution (that is, <10 μm) patterning or to fabricate devices in electronics or other areas of technology by use of functional or sacrificial inks. Here, we introduce methods and materials for e-jet printing with resolution well within the submicrometre range. Patterning of wide ranging classes of inks in diverse geometries illustrates some of the capabilities. Printed electrodes for functional transistors and representative circuit designs demonstrate potential applications in electronics. These results define some advantages and disadvantages of this approach, in its current form, compared with other ink printing techniques.

Figure 1 shows a schematic diagram of our e-jet printing system. A syringe pump (flow rates ≤ ~30 pl s^{-1}) or pneumatic pressure controller (applied pressure ≤ ~5 psi) connected to a

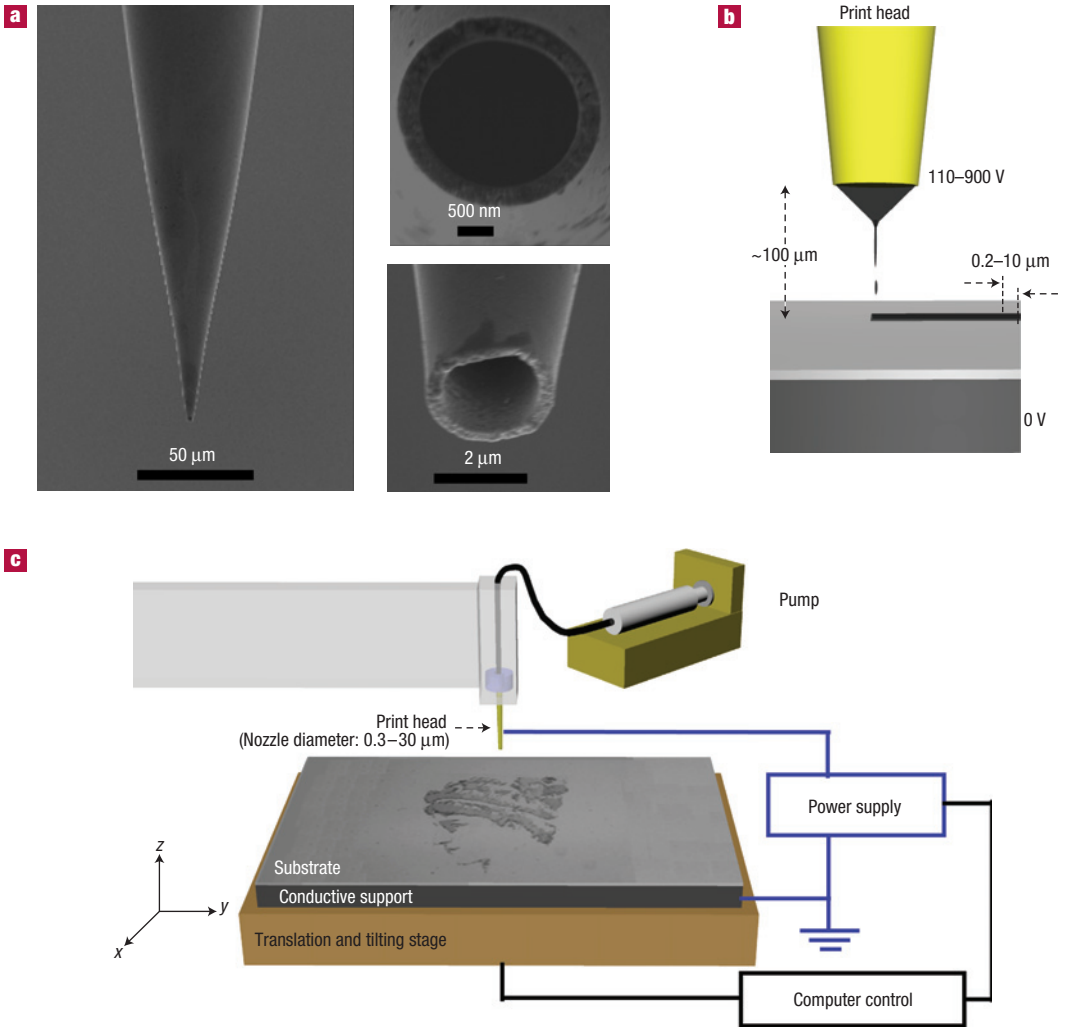


Figure 1 Nozzle structures and schematic diagrams of a high-resolution e-jet printer. a, SEM images of a gold-coated glass microcapillary nozzle (2 μm internal diameter). A thin film of surface-functionalized Au coats the entire outer surface of the nozzle as well as the interior near the tip. The insets on the right show views of this tip region. **b**, Nozzle and substrate configuration for printing. Ink ejects from the apex of the conical ink meniscus that forms at the tip of the nozzle owing to the action of a voltage applied between the tip and ink, and the underlying substrate. These droplets eject onto a moving substrate to produce printed patterns. In this diagram, the substrate motion is to the right. Printed lines with widths as small as 700 nm can be achieved in this fashion. **c**, Printer set-up. A gold-coated nozzle (internal diameter: 1, 2 or 30 μm) is located above a substrate that rests on a grounded electrode with a separation (H) of ~100 μm. A power supply connects to the nozzle and the electrode under the substrate. The substrate/electrode combination mounts on a five-axis (*x*, *y*, *z* axes and two tilting axes in the *x*–*y* plane) stage for printing.

glass capillary (internal diameter of between 0.3 and 30 μm and outer diameter of between 0.5 and 45 μm) delivers fluid inks to the cleaved end of the capillary, which serves as a nozzle. The nozzle fabrication process is described in the Methods section. Figure 1a shows scanning electron microscope (SEM) images of the nozzle and the nozzle opening. A thin film of sputter-deposited gold coats the entire outside of the microcapillary as well as the area around the nozzle and the inner surfaces near the tip. A hydrophobic self-assembled monolayer (1H,1H,2H,2H-perfluorodecane-1-thiol) formed on the gold limits the extent to which the inks wet the regions near the nozzle, thereby minimizing the probability for clogging and/or erratic printing behaviour (see Supplementary Information, Table S1). We refer to this functionalized gold-coated capillary mounted on a mechanical support fixture and connected to the pump as the e-jet print head. The nozzles used in these print heads have internal diameters that are much smaller than those used in previous work on e-jet printing[26–29], where the focus was on relatively low-resolution applications in graphic arts. The small nozzle dimensions are critically important to achieving high-resolution performance for device fabrication, for reasons described below.

A voltage applied between the nozzle and a conducting support substrate creates electrohydrodynamic phenomena that drive flow of fluid inks out of the nozzle and onto a target substrate. This substrate rests on a metal plate that provides an electrically grounded conducting support. The plate, in turn, rests on a plastic vacuum chuck that connects to a computer-controlled x, y and z axes translation stage. A two-axis tilting mount on top of the translation stage provides adjustments to ensure that motion in the x and y directions does not change the separation (H, typically ~100 μm) between the nozzle tip and the target substrate. A d.c. voltage (V) applied between the nozzle and the metal plate with

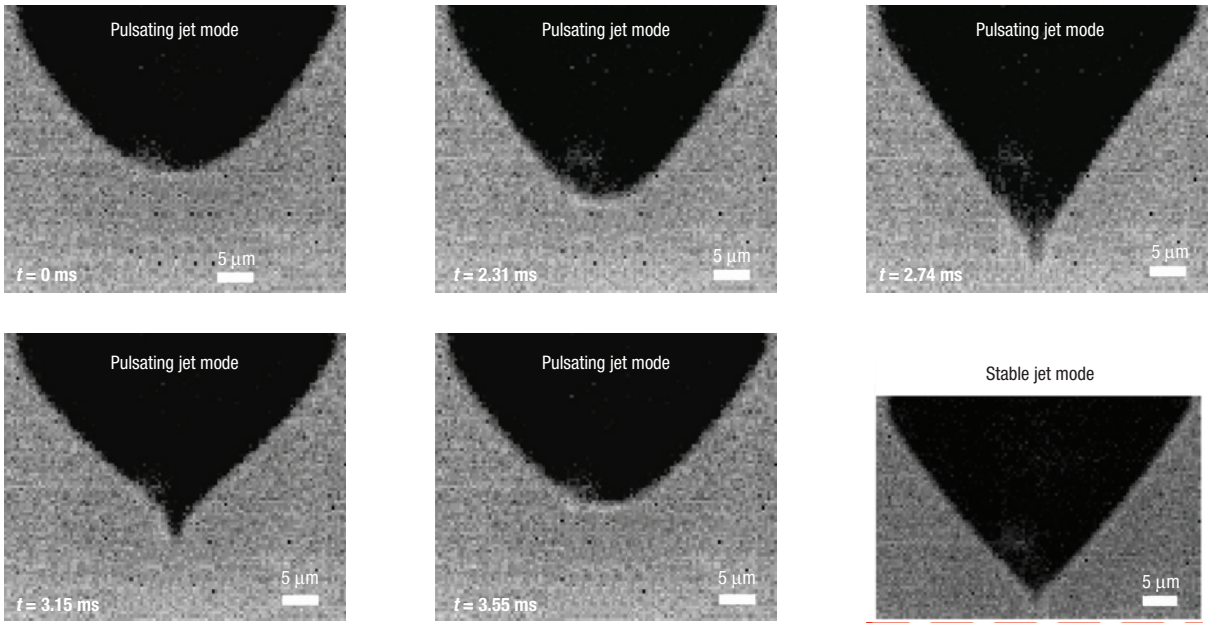


Figure 2 Time-lapse images of the pulsating liquid meniscus in one cycle at $V/H = 3.5\,\mathrm{V\,\mu m^{-1}}$. V is the applied voltage between the nozzle and the substrate and H is the distance between the nozzle tip and the substrate. The bottom right image corresponds to the stable jet mode, which is achieved at $V/H \sim 9\,\mathrm{V\,\mu m^{-1}}$ for this system. These images were captured at a frame rate of 66,000 fps and exposure time of 11 μs, using a high-speed camera. The reference time ($t = 0$) corresponds to the time at which the meniscus first reaches its fully retracted state.

a computer-controlled power supply generates an electric field that causes mobile ions in the ink to accumulate near the surface of the pendent meniscus at the nozzle. The mutual coulombic repulsion between these ions induces a tangential stress on the liquid surface, thereby deforming the meniscus into a conical shape, known as a Taylor cone[30]. At sufficiently high electric fields, this electrostatic (Maxwell) stress overcomes the capillary tension at the apex of the liquid cone; droplets eject from the apex to expel some portion of the surface charge (Rayleigh limit). Even very small ion concentrations are sufficient to enable this ejection process. For example, in uncontrolled spray modes, ejection is possible with liquids that have electrical conductivities that span ten decades[31], from 10^{-13} to $10^{-3}\,\mathrm{S\,m^{-1}}$. Coordinating the operation of the power supply with the system of translation stages enables direct-write e-jet printing of inks in arbitrary geometries (see Fig. 1b,c).

To understand the fundamental dynamics of this electric-field-driven jetting behaviour, a high-speed camera (Phantom 630, 66,000 fps) was used to image the process of Taylor-cone deformation and droplet ejection directly at the nozzle. For these experiments, an aqueous ink of a blend of poly(3,4-ethylenedioxythiophene) and poly(styrenesulphonate) (PEDOT/PSS) was used. The images, shown in Fig. 2, show that the meniscus at the nozzle expands and contracts periodically owing to the electric field. A complete cycle, which occurs in roughly 3–10 ms for the systems investigated here, consists of stages of liquid accumulation, cone formation, droplet ejection and relaxation[32]. The initial spherical meniscus at the nozzle tip changes gradually into a conical form owing to the accumulation of surface charges. The radius of curvature at the apex of the cone decreases until the Maxwell stress matches the maximum capillary stress, resulting in charged fluid jet ejection. This ejection decreases the cone volume and charges, thereby reducing the electrostatic stress to values less than the capillary tension. The ejection then stops and the meniscus retracts to its original spherical shape. The apex of the cone can oscillate, leading to the ejection of multiple droplets in short bursts. The frequency of this oscillation, which is in the kHz frequency range, increases in a nonlinear fashion with the electric field[33,34]. After a period of ejection in the form of multiple pulsations similar to the cycle shown in Fig. 2, the retracted spherical meniscus remains stable and largely unperturbed until the next period of ejection. This accumulation time depends on the flow rate imposed by the pump and on the electrical charging times associated with the resistance and capacitance of the system[33,34].

At sufficiently high fields, a stable jet mode (as opposed to the pulsating mode described above) can be achieved. In this situation, a continuous stream of liquid emerges from the nozzle, as shown in Fig. 2. At even higher fields, multiple jets can form, culminating ultimately in an atomization mode (e-spray mode) of the type used in mass spectroscopy and other well-established fields of application[35,36]. For controlled high-resolution printing of the type introduced here, this mode must be avoided. Either the stable jet or the pulsating modes can be used. The sensitivity of the stable jet mode to applied fields (too high results in uncontrolled spray, and too low results in pulsation) favours, in a practical sense, the pulsating operation. A key to achieving high resolution, from the standpoint of print-head design, is the use of fine nozzles with sharp tips. Such nozzles lead directly to small droplets/streams. In addition, the low V and H values that result from electric-field-line focusing at the sharp tips of such nozzles and the distribution of the electric field lines themselves combine to minimize lateral variations in the placement of the droplets/streams on the printed substrate (see Supplementary Information, Fig. S1).

A wide range of functional organic and inorganic inks, including suspensions of solid objects, can be printed using this approach, with resolutions extending to the submicrometre range. Figure 3a,b shows dot-matrix text patterns formed using a solution ink of a conducting polymer PEDOT/PSS and a photocurable polyurethane prepolymer (NOA 74, Norland Products) printed

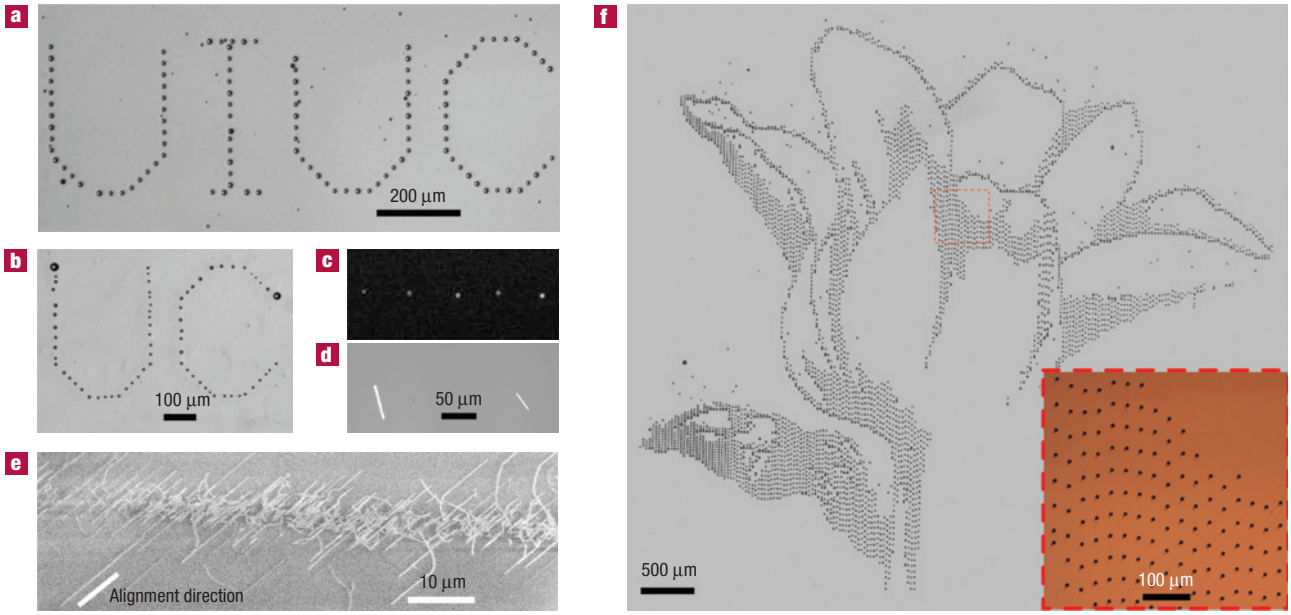


Figure 3 Optical micrographs and SEM images of various images formed with different inks. a, Letters printed with the conducting polymer PEDOT/PSS. The average dot diameter is 10 μm. **b**, Letters printed with a photocurable polyurethane polymer with dot diameters of 10 μm. **c**, Fluorescence optical micrograph (emission at 680 nm) of Si nanoparticles (average diameter of 3 nm) printed from a suspension in 1-octanol. The diameter of the printed dots is 4 μm. **d**, Optical micrograph of single-crystal Si rods (thickness: 3 μm, length: 50 μm, width: 2 μm) printed from a suspension in 1-octanol. **e**, SEM image of aligned SWNTs grown by CVD on quartz using printed patterns of ferritin as a catalyst. **f**, Image of a flower formed with printed dots (~8 μm diameters) of SWNTs from an aqueous solution. In all cases, nozzles with internal diameters of 30 μm were used.

onto a SiO_2(300 nm)/Si substrate. Figure 3c,d shows examples of printed inks that consist of suspensions of Si nanoparticles (average diameter: 3 nm)[37] and single-crystal Si rods (length: 50 μm, width: 2 μm, thickness: 3 μm)[38] dispersed in 1-octanol. The Si nanoparticles emit fluorescent light at 680 nm wavelength, as shown in Fig. 3c. Suspensions of ferritin nanoparticles can also be printed and then used as catalytic seeds for the chemical vapour deposition (CVD) growth of single-walled carbon nanotubes (SWNTs). Figure 3e shows the results, in which the printing and growth occurred on an annealed ST (stable temperature)-cut quartz substrate[39], to yield well-aligned individual SWNTs. For the structures printed onto SiO_2/Si, the silicon formed the conducting support for printing. In the case of quartz, a metal supporting plate was used. Computer-coordinated control of the power supply and the stages enables printing of complex patterns, such as digitized graphic images or circuit layouts. Figure 3f shows a printed image of a flower formed with an ink consisting of surfactant-stabilized SWNTs in water[40]. The average dot diameter is 8 ± 0.3 μm, and the uniformity in the sizes is shown in the Supplementary Information, Fig. S2. For the results in Fig. 3, the nozzle internal diameter was 30 μm and the substrates moved at speeds of ~100 μm s^{-1} (1 mm s^{-1} for Fig. 3a,b). These conditions yielded dot-matrix versions of the images with ~10 μm dot diameters. These dots are associated with the accumulation of multiple micro/nanodroplets ejected at the kHz level frequency in the pulsating mode; the separation between these dots corresponds to the accumulation time mentioned previously. (For Fig. 3d, owing to the low concentration of Si rods (~5 rods nl^{-1}), a relatively large drop diameter of ~100 μm was selected by applying the voltage for 100 ms with the nozzle held fixed.)

Although the ~10 μm feature sizes shown in Fig. 3 are suitable for various applications, the resolution can be improved by using smaller nozzles. Supplementary Information, Fig. S3 shows a portrait image composed of 2 μm dots printed with a 2-μm-internal-diameter nozzle and a printing speed of 20 μm s^{-1}. The printing resolution can be extended much further into the submicrometre regime. Figure 4a shows an image of the ancient scholar, Hypatia, printed using polyurethane ink. Dots ~490 nm in diameter were achieved with a 500-nm-internal-diameter nozzle for this case. Further reducing the internal diameter to 300 nm reduces the dot size to 240 ± 50 nm, as shown in Fig. 4b. Patterns of continuous lines and other shapes can be achieved by printing at stage translation speeds that allow the dots to merge. Figure 4c shows patterns of lines printed onto a SiO_2/Si substrate using the 2-μm-internal-diameter nozzle and a printing speed of 10 μm s^{-1}; the line widths, for single-pass printing, are ~3 μm. With a 1-μm-internal-diameter nozzle, line widths of ~700 nm can be achieved using polyethyleneglycol methyl ether solution (Aldrich), as shown in Fig. 4d. These results represent a resolution that significantly exceeds that of conventional unassisted thermal- or piezoelectric-type ink-jet systems. The slight 'waviness' in the submicrometre dots or lines in Fig. 4a,b,d is due to the combined effects of mechanical resonances in the long capillary used in the print head and slight fluctuations associated with the e-jet process.

Printed electronics represents an important application area that can take advantage of both the extremely high-resolution capabilities of e-jet printing as well as its compatibility with a range of functional inks. To demonstrate the suitability of e-jet printing for fabricating key device elements in printed electronics, we patterned complex electrode geometries for ring oscillators, source/drain electrodes for transistors, and we built working transistors. In these examples, a photocurable polyurethane precursor provided a printable resist layer for patterning metal electrodes by chemical etching. The print head in this case used a

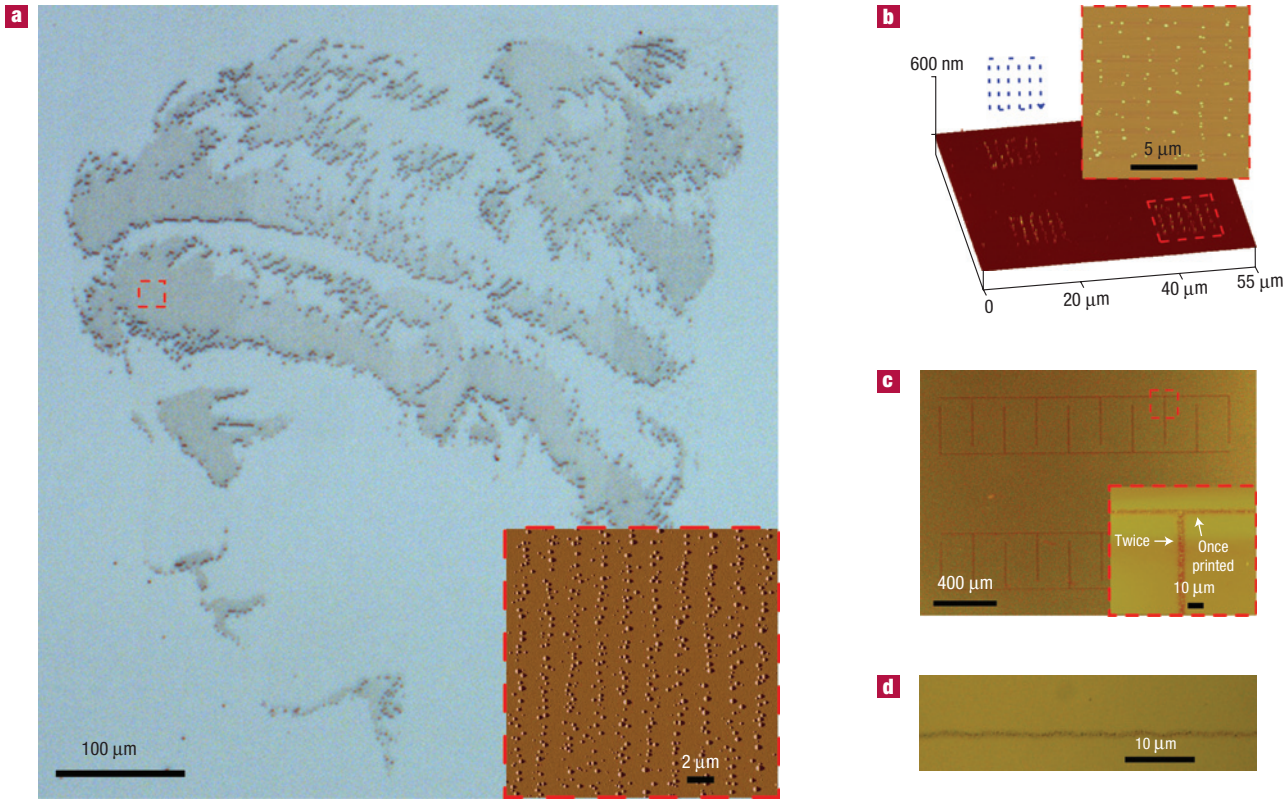


Figure 4 High-resolution e-jet printing with printed feature sizes in the range from ~240 nm to ~5 μm. a, Optical micrograph of a portrait of the ancient scholar, Hypatia, printed using a polyurethane ink and a 500-nm-internal-diameter nozzle. The diameters of the dots are ~490 nm. The inset shows an AFM image of the printed dots. **b**, Three-dimensional AFM image of aligned arrays of dots with diameters of 240 ± 50 nm, formed using the polyurethane and a 300-nm-internal-diameter nozzle. The blue dashed lines show the scan direction of the nozzle, and the top right inset shows a magnified AFM image of the printed dot array. **c**, Continuous lines printed using the SWNT ink and a 2-μm-internal-diameter nozzle. The horizontal lines (widths: ~3 μm) were printed in a single pass, whereas the vertical lines (width: ~5 μm) were formed by printing in two passes. **d**, Optical micrograph of a printed line of polyethyleneglycol (width: 700–800 nm) formed using a 1-μm-internal-diameter nozzle.

1-μm-internal-diameter nozzle; the printing speed was 100 μm s^{-1}. The substrate consisted of SiO_2 (300 nm)/Si coated uniformly with Au (130 nm) and Cr (2 nm). Figure 5a shows a pattern of printed polyurethane after curing by exposure to ultraviolet light (~1 J cm^{-2}). The resolution was 2 ± 0.4 μm, as defined by the minimum line widths. Much larger features, shown here in the form of electrode pads with dimensions up to 1 mm, are possible by overlapping the fine lines. Wet etching the printed substrate (Au etchant: trifluoroacetic acid, Transene; Cr etchant: Cr mask etchant, Transene) removed the Au/Cr bilayer in regions not protected by the polyurethane. Removing the polyurethane by soaking in methylene chloride and, in some cases, by oxygen plasma etching (plasmatherm reactive ion etch system, 20 s.c.c.m. O_2 flow with a chamber base pressure of 150 mtorr, 150 W, and radiofrequency power for 5 min), completed the fabrication or prepared the substrate for deposition of the next functional material. Figure 5b–e shows various patterns of Au/Cr electrodes formed in this manner. Figure 5d shows an array of printed source/drain electrodes with different spacings (that is, channel lengths, L). As shown in the inset of Fig. 5d, channel lengths as small as 1 ± 0.2 μm can be achieved with channel widths of up to hundreds of micrometres (~170 μm in this case). An atomic force microscopy (AFM) image of part of the channel area shows sharp, well-defined edges (Fig. 5e). The ability to print channel lengths with sizes in the micrometre range in a direct fashion, without the use of substrate wetting or relief assist features, is important owing to the key role of this dimension in determining the switching speeds and the output currents of the transistors.

As a demonstration of device fabrication by e-jet printing, thin-film transistors (TFTs) that use perfectly aligned arrays of SWNTs[41] as the semiconductor and e-jet-printed electrodes for source and drain were fabricated on flexible plastic substrates. The fabrication process began with e-beam evaporation of a uniform gate electrode (Cr: 2 nm/Au: 70 nm/Ti: 10 nm) onto a sheet of polyimide (thickness: 25 μm). A layer of SiO_2 (thickness: 300 nm) deposited by plasma-enhanced CVD at 250 °C and a spin-cast film of epoxy (SU-8, thickness: 200 nm) formed a bilayer gate dielectric. The epoxy also served as an adhesive for the dry transfer of SWNT arrays grown by CVD on quartz wafers using patterned stripes of iron catalyst[41]. Evaporating uniform layers of Cr (2 nm)/Au (100 nm) onto the transferred SWNT arrays, followed by e-jet printing and photocuring of polyurethane and then etching of the exposed parts of the Cr/Au to define source/drain electrodes completed the fabrication of devices with different channel lengths, L. SWNTs outside the channel areas were removed by reactive ion etching (150 mtorr, 20 s.c.c.m. O_2, 150 W, 30 s) to isolate these devices. Figure 6a,b shows schematic diagrams of the device layouts and an SEM image of the aligned SWNTs with the e-jet-printed source/drain electrodes. The arrays consist of ~2.5 SWNTs/10 μm. Figure 6c shows typical transfer characteristics that indicate the expected p-channel behaviour[42]. The current outputs increase approximately linearly with $1/L$, with ratios of the 'on' to the 'off'

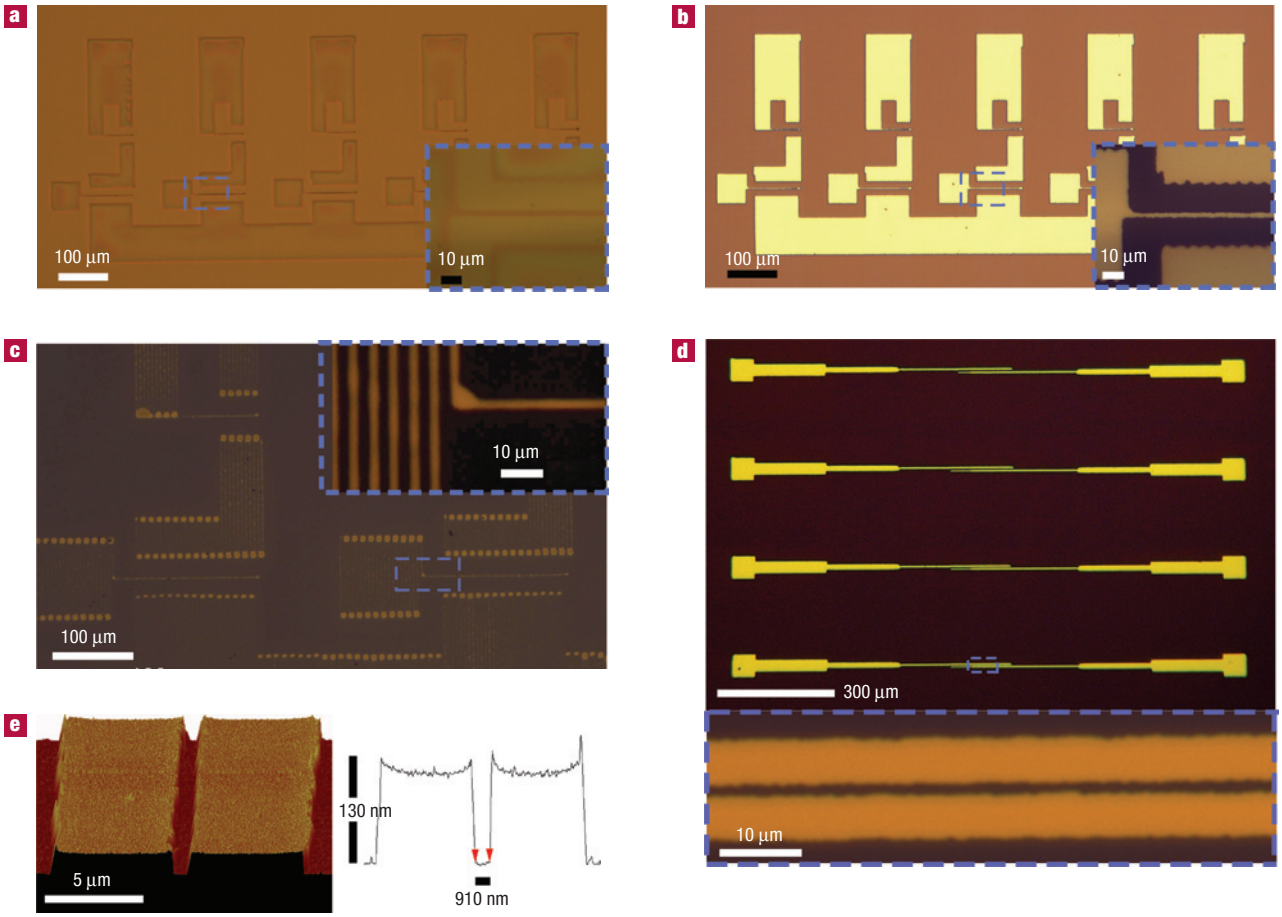


Figure 5 Patterns of electrode structures for a ring oscillator and isolated transistors formed by e-jet printing of a photocurable polyurethane ink that acts as an etch resist for a uniform underlying layer of metal (Au/Cr). a, E-jet-printed polyurethane etch resist for a ring oscillator circuit before etching the metal layers. **b**, Patterned Au electrode lines with ~2 μm width after etching and stripping the resist. The insets show magnified images. **c**, Au electrode lines (widths ~2 μm). **d**, Array of source/drain electrode pairs formed by e-jet printing of the resist layer, etching of metal and then stripping the resist. The inset shows an electrode pair separated by ~1 μm. **e**, AFM image and depth profile of a portion of this pair.

currents that are between ~1.5 and ~4.5 (inset of Fig. 6c), as expected owing to the population of metallic tubes in the arrays. Figure 6d (black circles) shows approximate device mobilities evaluated in the linear regime, calculated from the physical widths of source/drain electrodes ($W = 80\,\mu m$), a parallel-plate model for capacitance (C), and the transfer curves, according to $\mu_{dev} = (L/WCV_D) \cdot (\partial I_D/\partial V_G)$. These mobilities are between 7 and 42 $cm^2\,V^{-1}\,s^{-1}$ with L in the range of 1–42 μm, and decrease with L owing to the contact resistance[41–43]. The nature of electrostatic capacitance coupling of the gate to the SWNTs is important to the behaviour of these devices. Calculations of this gate capacitance are given in the Supplementary Information. This accurate capacitance model yields mobilities of 20–141 $cm^2\,V^{-1}\,s^{-1}$, as shown in Fig. 6d (red squares). We speculate that exposing the tubes to etchants for Cr/Au can induce defects, thereby resulting in lower mobilities than those reported previously with devices fabricated by other means[41]. The on/off ratios can be enhanced by an electrical breakdown process[41]. Transfer curves evaluated before and after this process are compared in Fig. 6e, for the case of a transistor with $L = 22\,\mu m$. The on/off ratio improves to >1,000 without substantial reduction in mobility (28 to 21 $cm^2\,V^{-1}\,s^{-1}$). Figure 6f shows full current–voltage characteristics before (inset) and after breakdown. Figure 6g shows an optical micrograph of a set of devices on a flexible sheet of polyimide, and Fig. 6h shows the normalized mobility and on/off ratio as a function of bending-induced strain[44] (ε). No significant change in the mobility or on/off ratio occurs for bending to radii of curvature (R_C) as small as 2 mm.

The advantages of this high-resolution form of e-jet printing over conventional ink-jet printing lie mainly in the high levels of resolution that can be obtained. We expect, in fact, that further reductions in the nozzle dimensions will enable resolution even deeper into the nanoscale. Achieving resolution at this level represents a topic of current work. The speeds for printing, using the particular systems described here, are relatively slow, although multiple-nozzle implementations, conceptually similar to those used in conventional ink-jet print heads, can reduce the printing duration. Key findings from analysis of throughput with conventional ink-jet printing systems also apply to the e-jet approach[45]. A main disadvantage of the e-jet approach is that the printed droplets have substantial charge that might lead to unwanted consequences in resolution and in device performance, particularly when used with electrically important layers such as gate dielectrics and semiconductor films. The

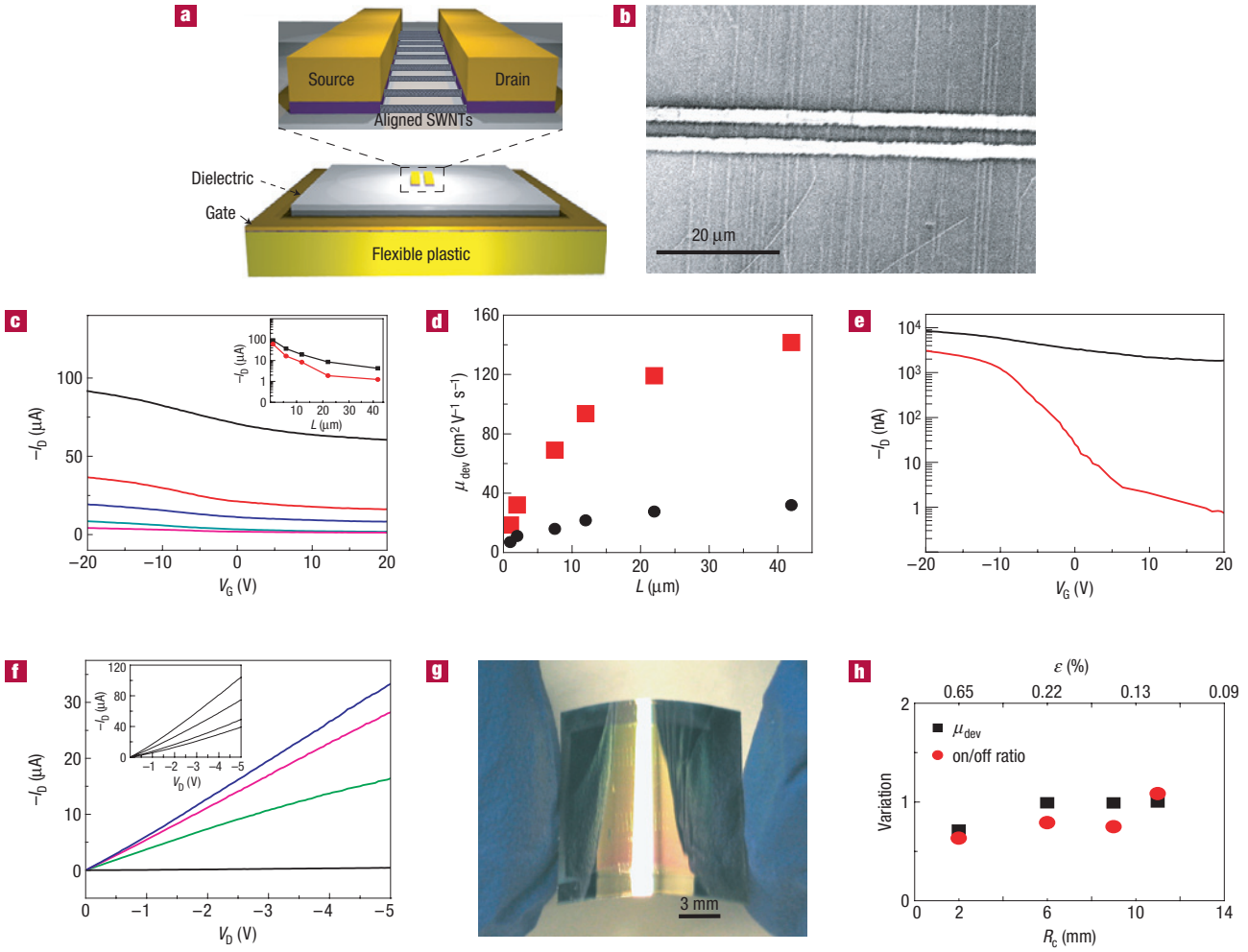


Figure 6 Fabrication of perfectly aligned SWNT-TFTs on a plastic substrate with e-jet printing for the critical features, that is, the source and drain electrodes. **a**, Schematic diagram of the transistor layout, where the source/drain electrodes are patterned by e-jet printing. **b**, SEM image of the aligned SWNTs connected by e-jet-printed source/drain electrodes. The tube density is ~2.5 SWNTs/10 μm. **c**, Transfer curves measured from transistors with channel lengths, $L = 1$, 6, 12, 22 and 42 μm, from top to bottom, and channel widths, $W = 80$ μm at a source/drain voltage, $V_D = -0.5$ V. The inset shows on and off currents (black and red lines, respectively) as a function of L. **d**, Linear-regime device mobilities (μ_{dev}) calculated from the parallel-plate (black circles) and rigorous (red squares) capacitance models, as a function of L. **e**, Transfer curves from a transistor with $L = 22$ μm before (black line) and after (red) an electrical breakdown process. This breakdown reduces the 'off' current to less than ~1 nA to yield an on/off ratio of ~1,000. **f**, Current–voltage characteristics recorded after the electrical breakdown process. The gate voltage varies between −20 and 10 V in steps of −10 V, from top to bottom. The inset shows the current–voltage curve before the breakdown with the same gate voltages for comparison. **g**, Photograph of an array of flexible SWNT-TFTs. **h**, Variation of the normalized mobility (black squares) and on/off ratio (red circles) of a SWNT-TFT as a function of bending-induced strain (ε) and the radii to curvature (R_c).

effects of this charge might be minimized by using high-frequency alternating driving voltages for the e-jet process. These and other process improvements, together with exploration of applications in biotechnology and other areas, represent promising areas for future work.

METHODS

PREPARATION OF NOZZLES

Au/Pd (70 nm thickness) and Au (50 nm) layers were coated onto glass pipettes with tip internal diameters of between 0.3 and 30 μm (World Precision Instruments) using a sputter coater (Denton, Desk II TSC). Dipping the tip of the metal-coated pipette into 1H, 1H, 2H, 2H-perfluorodecane-1-thiol (Fluorous Technologies) solution (0.1 wt% in dimethylformamide) for 10 min formed a hydrophobic self-assembled layer on the gold surface of the nozzle tip. The capillary was connected to a syringe pump (Harvard Apparatus, Picoplus) or a pneumatic pressure actuator through a polyethylene tube.

SYNTHESIS OF FUNCTIONAL INKS

PEDOT/PSS ink: PEDOT/PSS (Baytron P, H.C. Starck) was diluted with H_2O (50 wt%), and then mixed with polyethyleneglycol methyl ether (Aldrich, 15 wt%) to reduce the surface tension (to lower the voltage needed to initiate printing) and the drying rate at the nozzle.

Single-crystal Si rods: Patterning the top Si layer (thickness: ~3 μm) of a silicon-on-insulator wafer by reactive ion etching, and then etching the underlying SiO_2 with an aqueous etchant of HF (49%)[38] with 0.1% of a surfactant (Triton X-100, Aldrich) formed the rods. These rods were suspended

in H_2O and then filtered through filter paper (pore size: 300 nm). The rods were then suspended in 1-octanol. After printing this ink, the surfactant residue was thermally removed by heating to 400 °C in air for 5 h.

Ferritin: First, ferritin (Sigma) was diluted in H_2O with a volume ratio of 1(ferritin):200(H_2O). Then 1 wt% of a surfactant (Triton X-100) was added to this solution to reduce the surface tension (to lower the voltage needed to initiate printing). The surfactant residue was removed at 500 °C before CVD growth of SWNTs.

SWNT solution: Single-walled carbon nanotubes produced by the electric arc method (P2-SWNT, Carbon Solution) were suspended in aqueous octyl-phenoxy-polyethoxyethanol (Triton X-405, 2 wt%). The concentration was ~6.9 mg l^{-1}.

PREPARATION OF SUBSTRATES

Si wafers with 300-nm-thick layers of thermal SiO_2 (Process Specialties) were used as substrates. The underlying Si was electrically grounded during printing. A glass slide (thickness: ~100 μm) was used for fluorescence optical micrography (Fig. 3c), and an ST-cut quartz wafer was used after annealing at 900 °C for guided growth of SWNTs (Fig. 3e). Here, the glass/quartz substrates were placed on an electrically grounded metal plate during printing. For printing of complex images (Figs 3f, 4a,b), the Si wafers were exposed to perfluorosilane vapour before e-jet printing to produce a hydrophobic self-assembled monolayer.

Received 23 February 2007; accepted 21 June 2007; published 5 August 2007.

References

1. Forrest, S. R. The path to ubiquitous and low-cost organic electronic applications on plastics. *Nature* **428**, 911–918 (2004).
2. Gans, B. J., Duineveld, P. C. & Schubert, U. S. Inkjet printing of polymers: State of the art and future development. *Adv. Mater.* **16**, 203–213 (2004).
3. Parashkov, R., Becker, E., Riedl, T., Johannes, H. & Kowalsky, W. Large area electronics using printing method. *Proc. IEEE* **93**, 1321–1329 (2005).
4. Chang, P. C. *et al.* Film morphology and thin film transistor performance of solution-processed oligothiophenes. *Chem. Mater.* **16**, 4783–4789 (2004).
5. Sirringhaus, H. *et al.* High-resolution inkjet printing of all-polymer transistor circuits. *Science* **290**, 2123–2126 (2000).
6. Shimoda, T. *et al.* Solution-processed silicon films and transistors. *Nature* **440**, 783–786 (2006).
7. Burns, S. E., Cain, P., Mills, J., Wang, J. & Sirringhaus, H. Inkjet printing of polymer thin-film transistor circuits. *Mater. Res. Soc. Bull.* **28**, 829–834 (2003).
8. Wong, W. S., Ready, S. E., Lu, J. P. & Street, R. A. Hydrogenated amorphous silicon thin-film transistor arrays fabricated by digital lithography. *IEEE Electron Device Lett.* **24**, 577–579 (2003).
9. Szczech, J. B., Megaridis, C. M., Gamota, D. R. & Zhang, J. Fine-line conductor manufacturing using drop-on-demand PZT printing technology. *IEEE Trans. Electron. Packag. Manufactur.* **25**, 26–33 (2002).
10. Shimoda, T., Morii, K., Seki, S. & Kiguchi, H. Inkjet printing of light-emitting polymer displays. *Mater. Res. Soc. Bull.* **28**, 821–827 (2003).
11. Chang, S. C. *et al.* Multicolor organic light-emitting diodes processed by hybrid inkjet printing. *Adv. Mater.* **11**, 734–737 (1999).
12. Hebner, T. R. & Sturm, J. C. Local tuning of organic light-emitting diode color by dye droplet application. *Appl. Phys. Lett.* **73**, 1775–1777 (1998).
13. Lemmo, A. V., Rose, D. J. & Tisone, T. C. Inkjet dispensing technology: Application in drug discovery. *Curr. Opin. Biotechnol.* **9**, 615–617 (1998).
14. Heller, M. J. DNA microarray technology: Devices, systems, and applications. *Annu. Rev. Biomed. Eng.* **4**, 129–153 (2002).
15. Nallani, A., Chen, T., Lee, J. B., Hayes, D. & Wallace, D. Wafer level optoelectronic device packaging using MEMS. *Proc. SPIE: Smart Sensors Actuators MEMS II* **5836**, 116–127 (2005).
16. Bietsch, A., Zhang, J., Hegner, M., Lang, H. P. & Gerber, C. Rapid functionalization of cantilever array sensors by inkjet printing. *Nanotechnology* **15**, 873–880 (2004).
17. Hiller, J., Mendelsohn, J. D. & Rubner, M. F. Reversibly erasable nanoporous anti-reflection coatings from polyelectrolyte multilayers. *Nature Mater.* **1**, 59–63 (2002).
18. Ling, M. M. & Bao, Z. Thin film deposition, patterning, and printing in organic thin film transistors. *Chem. Mater.* **16**, 4824–4840 (2004).
19. Calvert, P. Inkjet printing for materials and devices. *Chem. Mater.* **13**, 3299–3305 (2001).
20. Sanaur, S., Whalley, A., Alameddine, B., Carnes, M. & Nuckolls, C. Jet-printed electrodes and semiconducting oligomers for elaboration of organic thin-film transistors. *Org. Electron.* **7**, 423–427 (2006).
21. Cheng, K. *et al.* Inkjet printing, self-assembled polyelectrolytes, and electroless plating: Low cost fabrication of circuits on a flexible substrate at room temperature. *Macromol. Rapid Commun.* **26**, 247–264 (2005).
22. Creagh, L. T. & McDonald, M. Design and performance of inkjet printheads for non graphic arts applications. *Mater. Res. Soc. Bull.* **28**, 807 (2003).
23. Wang, J. Z., Gu, J., Zenhausern, F. & Sirringhaus, H. Low-cost fabrication of submicron all polymer field effect transistors. *Appl. Phys. Lett.* **88**, 133502 (2006).
24. Stutzmann, N., Friend, R. H. & Sirringhaus, H. Self-aligned, vertical channel, polymer field effect transistors. *Science* **299**, 1881–1885 (2003).
25. Sele, C. W., Werne, T., Friend, R. H. & Sirringhaus, H. Lithography-free, self-aligned inkjet printing with sub-hundred nanometer resolution. *Adv. Mater.* **17**, 997–1001 (2005).
26. Mills, R. S. *Recent Progress in Ink Jet Technologies II* 286–290 (Society for Imaging Science and Technology, Washington, 1999).
27. Nakao, H., Murakami, T., Hirahara, S., Nagato, H. & Nomura, Y. *IS&Ts NIP15: 1999 International Conference on Digital Printing Technologies* 319–322 (Society for Imaging Science and Technology, Washington, 1999).
28. Choi, D. H. & Lee, F. C. *Proc. of IS&T's Ninth International Congress on Advances in Non-Impact Printing Technologies. October 4–8, Yokohama, Japan* (Society for Imaging Science and Technology, Washington, 1993).
29. Kawamoto, H., Umezu, S. & Koizumi, R. Fundamental investigation on electrostatic ink jet phenomena in pin-to-plate discharge system. *J. Imaging Sci. Technol.* **49**, 19–27 (2005).
30. Taylor, G. Disintegration of water droplets in an electric field. *Proc. R. Soc. Lond. A* **280**, 383–397 (1964).
31. Jayasinghe, S. N. & Edirisinghe, M. J. Electric-field driven jetting from dielectric liquids. *Appl. Phys. Lett.* **85**, 4243 (2004).
32. Marginean, I., Parvin, L., Heffernan, L. & Vertes, A. Flexing the electrified meniscus: The birth of a jet in electrosprays. *Anal. Chem.* **76**, 4202–4207 (2004).
33. Chen, C. H., Saville, D. A. & Aksay, I. A. Scaling law for pulsed electrohydrodynamic drop formation. *Appl. Phys. Lett.* **89**, 124103 (2006).
34. Hayati, I., Bailey, A. I. & Tadros, T. F. Investigations into mechanisms of electrohydrodynamic spraying of liquids. *J. Colloid Interface Sci.* **117**, 205–221 (1987).
35. Wickware, P. & Smaglik, P. Mass spectroscopy: Mix and match. *Nature* **413**, 869 (2001).
36. Salata, O. V. Tools of nanotechnology: Electrospray. *Curr. Nanosci.* **1**, 25–33 (2005).
37. Smith, A. *et al.* Observation of strong direct-like oscillator strength in the photoluminescence of Si nanoparticles. *Phys. Rev. B* **72**, 205307 (2005).
38. Menard, E., Lee, K. J., Khang, D. Y., Nuzzo, R. G. & Rogers, J. A. A printable form of silicon for high performance thin film transistors on plastic substrates. *Appl. Phys. Lett.* **84**, 5398 (2004).
39. Kocabas, C., Shim, M. & Rogers, J. A. Spatially selective guided growth of high-coverage arrays and random networks of single-walled carbon nanotubes and their integration into electronic devices. *J. Am. Chem. Soc.* **128**, 4540–4541 (2006).
40. Park, J. U. *et al.* In situ deposition and patterning of single walled carbon nanotubes by laminar flow and controlled flocculation in microfluidic channels. *Angew. Chem. Int. Edn* **45**, 581–585 (2006).
41. Kang, S. J. *et al.* High performance electronics using dense, perfectly aligned arrays of single walled carbon nanotubes. *Nature Nanotechnol.* **2**, 230–236 (2007).
42. Chen, Z., Appenzeller, J., Knoch, J., Lin, Y. M. & Avouris, P. The role of metal-nanotube contact in the performance of carbon nanotube field effect transistors. *Nano Lett.* **5**, 1497–1502 (2005).
43. Kim, W. *et al.* Electrical contacts to carbon nanotubes down to 1 nm in diameter. *Appl. Phys. Lett.* **87**, 173101 (2005).
44. Lee, K. J. *et al.* A printable form of single-crystalline gallium nitride for flexible optoelectronic systems. *Small* **1**, 1164–1168 (2005).
45. Sheats, J. R. Manufacturing and commercialization issues in organic electronics. *J. Mater. Res.* **19**, 1974–1989 (2004).

Acknowledgements

The authors thank L. Jang and M. Nayfeh for supplying Si nanoparticle solutions, R. Shepherd and J. Lewis for the use of their high-speed camera, and R. Lin for assistance with setting initial experimental conditions. In addition, the authors acknowledge the Center for Nanoscale Chemical Electrical Mechanical Manufacturing Systems in the University of Illinois, which is funded by the National Science Foundation under grant DMI-0328162, and the Center for Microanalysis of Materials in University of Illinois, which is partially supported by the US Department of Energy under grant DEFG02-91-ER45439.

Correspondence and requests for materials should be addressed to J.A.R.

Supplementary Information accompanies this paper on www.nature.com/naturematerials.

Author contributions

J.-U.P. and J.A.R. designed the experiments and wrote the paper. J.-U.P. carried out the nozzle fabrication, ink design, printing and characterization. S.J.K. and J.-U.P. contributed to device fabrication. K.B., K.A., D.K.M., A.G.A. and P.M.F. designed the printing machine and contributed to project planning. J.G.G. was responsible for hydrodynamics analysis and project planning. C.Y.L. and M.S.S. synthesized SWNT solutions. M.H. developed the software algorithm and measured contact angles.

Competing financial interests

The authors declare no competing financial interests.

8

PATTERNING WITH ELECTROLYTE: SOLID-STATE SUPERIONIC STAMPING

K. H. Hsu, P. L. Schultz, N. X. Fang, and P. M. Ferreira

8.1 INTRODUCTION

Metallic structures are ubiquitous in micro- and nanotechnology. From interconnects in electronics to electrodes in sensors, batteries, and fuel cells, they play a pivotal role in the performance of devices [1–4]. Such metallic structures are also becoming increasingly important in emerging fields related to subwavelength optics, such as surface plasmons and plasmonic waveguides [5, 6]. Because metallic structures are such an integral part of ever-shrinking micro- and nanoscale devices and systems, it is of critical importance to be able to economically manufacture them at these length scales. However, the common practice for generating metallic patterns has relied on an indirect approach. For example, micro- or nanoscale patterns are first lithographically created in photoresist that is used as a sacrificial mold. Metal is deposited by evaporation or sputtering and the subsequent liftoff of the polymer and excess metal leaves behind the metal pattern [1]. Similarly, in the damascene process [7], pursued by the semiconductor industry, copper interconnects are created by electrochemically depositing copper into trenches patterned in a dielectric film. The inlaid metallic interconnect patterns are left behind after chemical–mechanical polishing is used to remove excess metal. Such indirect processes tend to be

Q1

Unconventional Nanopatterning Techniques and Applications by Hong H. Lee and John A. Rogers

expensive and complex, and require multiple process steps (around 20 per layer [8]) with stringent process environment control and very costly equipment.

There has been a renewed interest in direct patterning methods for metals in the electronics industry because of the high resistivity encountered in conductors with shrinking lateral dimensions. As the lateral dimensions approach the mean free electron path, roughly around 40 nm at 25°C, the electrical resistivity increases rapidly [9] (e.g., a 50-nm wire has twice the bulk resistivity of copper [10]). Grain boundary and sidewall scattering are thought to be the primary reasons. In the damascene process, the narrow trench geometries and impurities are thought to impede the grain growth rendering annealing and other grain growth techniques ineffective. As a result, process sequences in which blanket films are presented for patterning are thought to be advantageous [9]. Direct patterning such as metallic, primarily copper films, using reactive ion etching (RIE) or dry-etch processes, has been attempted by a number of researchers. Steinbruchel [11] gives a comprehensive discussion of the issues involved in RIE patterning of copper. Schwartz and Schaible [12] have used a CCl_4 while Ohno et al. [13, 14] have used a $SiCl_4$ mixture with N_2, Cl_2, and NH_3 and others [15] have used with BCl_3. In these processes, the high substrate temperatures (typically in the range of 250°C) make finding suitable mask materials a challenge. Low temperature etching of such metals is difficult because of the low volatility of the reaction products. Lee and Kuo [16] report the removal of $CuCl_2$ after a Cl_2 plasma etch with a dilute solution of HCl. Recent work by Tamirisa et al. [9] reports successful etching of copper at room temperatures by using alternating cycles of H_2 and Cl_2 plasmas; however, patterning results are yet to be reported. In summary, in spite of a need for direct, dry patterning of metal films such as Cu, the use of RIE processes has proved to be challenging.

Less conventional methods such as nanotransfer printing use a poly(dimethylsiloxane) (PDMS) stamp to directly place the pattern on the substrate and are suitable for micrometer-sized features; but often require pretreatment of the substrate [3]. Still other attempts have been made at using PDMS-based microcontact printing to deposit an etch resist prior to electrochemical etching of unwanted portions of film [17]. While economical, such techniques are primarily used for copper patterns with critical features between 1 and 500 μm. Nanosphere lithography [18] can be used for creating periodic patterns of metallic structures by using closely packed polystyrene microspheres as a mask. Another method, electrochemical micromachining, using a liquid electrolyte, has been proposed to directly produce submicrometer metallic features [19, 20]. Variations in resistance to flows and the high diffusion lengths of reacting species pose significant limitations on lateral extension of the features associated with the process. Accelerated etching at sharp edges and corners also leads to low geometrical fidelity in the pattern transferred from the electrochemical tool to substrate surface [19]. In addition, the use of liquids might contaminate both the tool and the substrate.

Given widespread use and need for metallic nanostructures, there is a general death of process technology for producing them directly. Therefore, it is of considerable interest to explore approaches that add to our repertoire of metallization

processes that are capable of directly and efficiently creating metallic structures with general geometries and nanometer-scale resolution. Q5

8.2 SOLID-STATE SUPERIONIC STAMPING

Solid-State Superionic Stamping (S4) [21, 22] is a solid-state electrochemical imprinting process that directly creates high resolution metallic nanopatterns in a single step. Similar to imprint lithography [23, 24], but different in that it does not pattern a polymer or use mechanical forces to squeeze it into a pattern, S4 directly imprints metals with an electrochemical reaction. Conceptually, it combines the large area, in-parallel, single-step process economics of nanoimprint lithography with the efficiency, precision, and low mechanical forces of electrochemical machining, as shown in Figure 8.1. At the center of this process is a solid electrolyte or superionic conductor [21]. Widely used in battery and fuel cell applications, due to their excellent ionic conductivity at room and relatively low temperatures, such materials offer the possibility of precise and efficient control of mass (ion) transport from

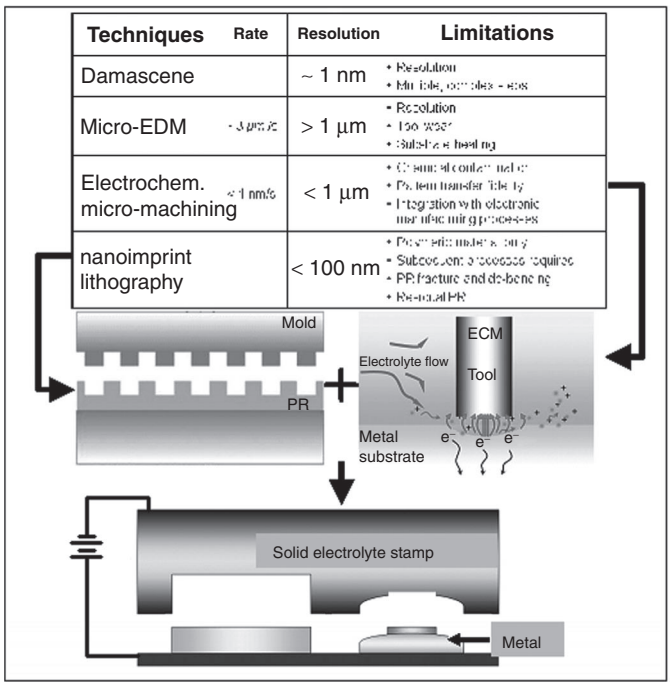


Figure 8.1. The solid-state superionic stamping processes combine the ideas of electrochemical machining and nanoimprint lithography, substituting PR for a metal film and liquid electrolyte by a nanopatterned solid electrolyte to produce a method for directly nanoimprinting metals.

or to a substrate to be patterned. A stamp made of a superionic conductor with a mobile cation (silver or copper sulfide, for example, in which the silver or copper ions are mobile) is prepatterned with fine features and brought into contact with the metallic substrate to be patterned. On the application of an electrical bias with the substrate as anode and a metallic electrode at the back of the stamp as cathode, a solid-state electrochemical reaction resulting in anodic dissolution of the metal begins at the contact interface with the stamp. At the anode–electrolyte interface, an appreciable potential drop causes the oxidation of metal atoms on the substrate to produce mobile cations. These mobilized ions migrate across the interface and through the interstitial channels and defect network in the lattice of the superionic conductor toward the cathode, until they recombine with electrons. The anodic dissolution progressively removes a metallic layer of the substrate at the contact area with the stamp. Assisted by a nominal pressure to maintain electrical contact, the stamp progresses into the substrate and generates a shape in the metallic substrate complementary to the prepatterned features on it. This idea is shown schematically in Figure 8.2. The advantage of using solid-state superionic conductors is that mass transport is restricted to the physical contact interface between the patterned electrolyte and the substrate (the anode), making it an ideal tool for nanoscale pattern transfer with high fidelity.

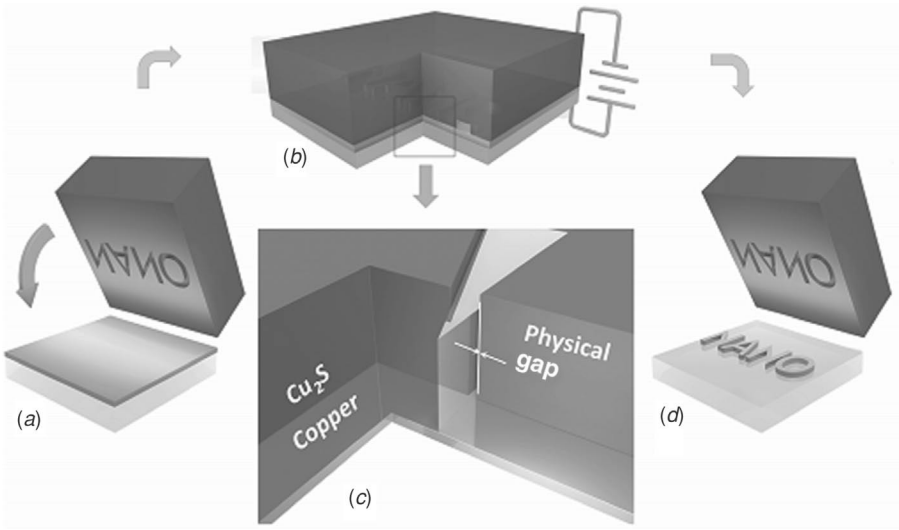


Figure 8.2. Schematic of the solid-state superionic stamping process [22]. (*a*) A prepatterned superionic conductor. In this case, Cu_2S is placed in contact with the substrate to be patterned. (*b*) An applied voltage with substrate as anode serves to initiate and propagate the anodic dissolution at the stamp–work piece interface. (*c*) At the interface, copper atoms are split into ions and electrons, which migrate through the crystal lattice and external circuit, respectively. (*d*) Removal of the stamp reveals the complementary pattern left behind. (Reproduced with permission from [22]. Copyright 2007 American Vacuum Society.)

The mobility of ions in an array of solid electrolytes, such as Ag_2S and $RbAg_4I_5$, has been exploited to create nanostructures in "direct-write" like processes. Sub-100-nm line and dot patterns have been written using an STM (scanning tunneling microscope) or AFM (atomic force microscope) tip [25, 26]. These techniques use Q2 an electric potential applied across a scanning probe and a solid electrolyte substrate surface to induce the migration of metal ions from within the substrate to the vicinity of the probe to form metallic clusters that create lines or dots on the electrolyte surface [27]. The practicality of this direct pattern writing as a manufacturing process is limited because of low throughput, difficulties in dimensional control of the structures formed, and processing parameters (the standoff distance of the probe must be precisely regulated and its travel speed must be coordinated to the growth of the structure. Most importantly, the metal structures are embedded on the electrolyte surface layer, making their subsequent use in any applications difficult. The S4 process, unlike these processes, is an area-patterning process. It is a subtractive process and produces the metallic pattern removing material from the substrate. Finally, the resulting nanopattern may reside on any substrate on which a metallic film can be deposited.

The introduction of soft or imprint lithography techniques [23, 24] for nanopatterning has attracted widespread research and commercial attention because of its ability to pattern large areas at resolutions reaching the single-digit nanometer range. The S4 process is an electrochemical imprinting process and like other imprint lithography processes uses a stamp or mold that is brought into contact with a substrate on which an imprint is to be generated. As a result it is very compatible with existing imprint equipment and adds to the repertoire of imprint lithography technology by providing a means of directly patterning metals without the need for imprinting polymer masks or molds.

To date, we have reported and demonstrated the use of this technique to create copper and silver nanostructures, reaching resolution as low as 30 nm with patterning rates (rate at which the stamp travels into the substrate) in the range of 0.1–5 nm s^{-1}. Figure 8.3 shows figure of a stamp as well as the results of the stamping process. The figure also draws correspondence with the process schematic. In the sections that follow, we will describe some of the salient features of this technology; manufacture of S4 stamps; the experimental stamping setup; monitoring the progress of the imprinting process; typical results obtained in using this process. We conclude with applications and directions for further development.

8.3 PROCESS TECHNOLOGY

Central to the S4 process is the stamp. As previously mentioned, the stamp material is a superionic conductor in which the metal ions are the mobile species. Generally, one can find candidates for electrochemical stamps from four categories of superionic solids: crystalline, amorphous glassy, polymeric, and composite [28]. Crystalline ionic conductors are further subdivided into soft- and hard-framework crystals, differing in their bond type (ionic or covalent), polarizability (high or low), and

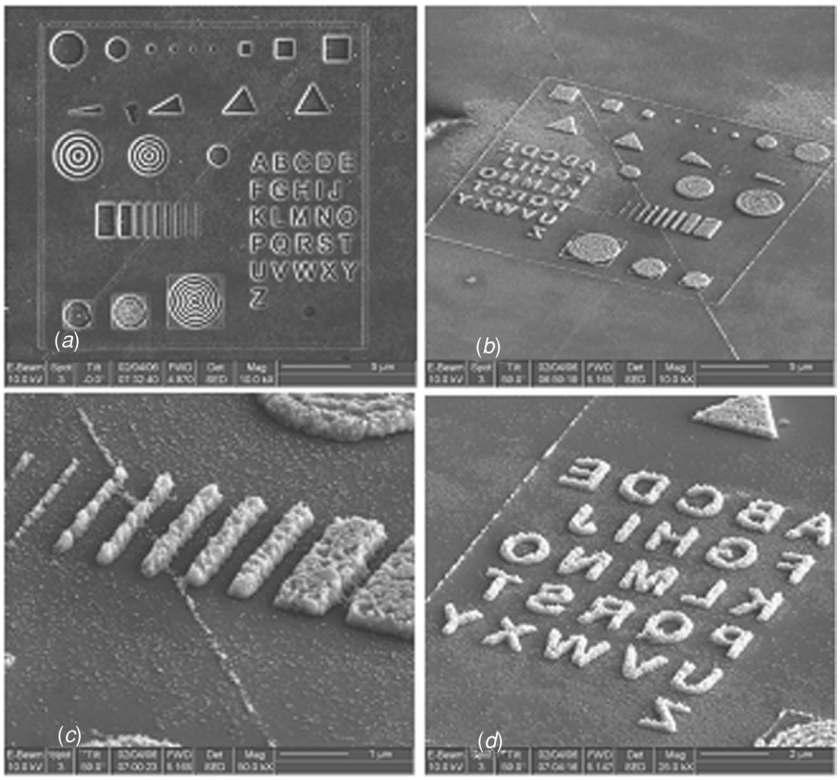


Figure 8.3. (*a*) A stamp made of silver sulfide and patterned using FIB with nanoscale alphabet and a series of nanoantennae patterns. (*b*) The results of superionic stamping. (*c*), (*d*) Isometric closeups of some of the patterning results. (Reproduced with permission from [21]. Copyright 2007 American Chemical Society.)

presence of a sharp order–disorder phase transition. While the amorphous-glassy type has high, isotropic conductivity and the absence of grain boundaries, polymeric electrolytes are characterized by combining polar polymers with ionic salts to achieve reduced weight and flexibility at the cost of lower ionic conductivity. Composite electrolytes result from the combination of several materials to improve the ionic conductivity at room temperature.

For the S4 process, it is necessary to choose stamp materials that are chemically stable and display high ionic conductivity at room or reasonably low (around 100°C) temperatures. Further, the stamp materials should be dense rather than porous and when polycrystalline should be relatively isotropic in terms of ionic conductivity. Finally, when possible, a ductile and malleable material facilitates ease of stamp preparation. The solid electrolyte system for silver is very rich, while that for copper, especially for room temperature ionic conductors, is quite restrictive. We chose Ag_2S and Cu_2S because of their stability, reasonably good mechanical and electrical

properties, and the relative ease of synthesis. Copper sulfide (Cu_2S), chalcocite, is a known mixed conductor wherein both electron holes and copper ions can carry charge. While its electronic conduction resembles that of a p-type semiconductor, its phase-dependent ionic conduction changes dramatically with temperature [29, 30]. In the γ-phase (orthorhombic) below 105°C, Cu_2S exhibits little or no ionic activity. At a transition temperature of around 105°C, its crystal structure changes from orthorhombic to hexagonal (β-phase), along with a significant drop in electronic conductivity and a dramatic increase, by a factor of 10^4, in ionic conductivity ($10^{-5}\,\Omega^{-1}\,cm^{-1}$ at 15°C to $10^{-1}\,\Omega^{-1}\,cm^{-1}$ at 105°C) [30]. Finally, at 470°C it undergoes another transition from β to α (cubic) in which the ionic conductivity drops dramatically again [31]. This variability of ionic conductivity can be exploited to tune the behavior of the S4 process by changing the operating temperature. Similarly, Ag_2S is a mixed ionic conductor with Ag^+ ions carrying charge. In the low temperature β-phase (acanthite) below 177°C, the ionic conductivity is a significant portion of the total conductivity and follows the Arrhenius type of relation with temperature, with a hopping activation energy of 0.2 eV [32, 33]. Following a transition to α-phase (argentite) above 177°C, the conductivity is predominantly electronic with a smaller dependence on temperature. In both phases, the conductivity is strongly influenced by composition, increasing with higher silver concentration. A large collection of ionic conductors with their ionic conductivities has been compiled by Agrawal and Gupta [28], should the reader wish to explore other stamp materials.

The silver or copper sulfide stamps are prepared by first synthesizing a dense silver or copper sulfide pellet in an improvized furnace with programmable temperature control. Sulfur powder with 99.999% purity (from Fisher Scientific Company) is hand pressed into a pellet of diameter 3 mm and inserted into a glass tube (3 mm internal diameter) with a silver or copper pellet (from Kurt J. Lesker Company). The two are then held together with a small constant force (required for obtaining a dense pellet) provided by a spring. The assembly is placed in the furnace at 400°C for 10 h to allow formation and annealing of sliver/copper sulfide. The pellets created in this manner are characterized with x-ray diffraction (Rigaku D-Max System with a scanning range (2θ) from 0° to 60° and a scan rate of $0.8^\circ\,min^{-1}$) and compared with standard peaks for the powder forms of β-silver or copper sulfide. Characterization of a number of pellets has confirmed the composition of the silver and copper sulfide pellets and the consistent output of the above process. Following growth, the pellets are machined to produce a conical end with a flat stamping surface (600 μm in diameter) that are either polished using lapping films of 1, 0.3, and 0.01 μm particle sizes, or trimmed with an ultramicrotome and a Diatome diamond knife (Leica EM UC6).

Patterning of the stamp is accomplished by both, focused ion beam (FIB) milling as well as embossing (in case of silver sulfide stamps). The patterns on the stamp, such as those shown in Figure 8.4*a*, were produced by FIB milling (FEI Dual-Beam DB-235) with a 50-pA aperture at a milling rate of about $50\,nm\,min^{-1}$. The deepest trench on the stamp in Figure 8.4*a* was about 250 nm. We have demonstrated direct embossing of the silver sulfide stamps against silicon masters such as calibration grids as shown in Figure 8.4*b*.

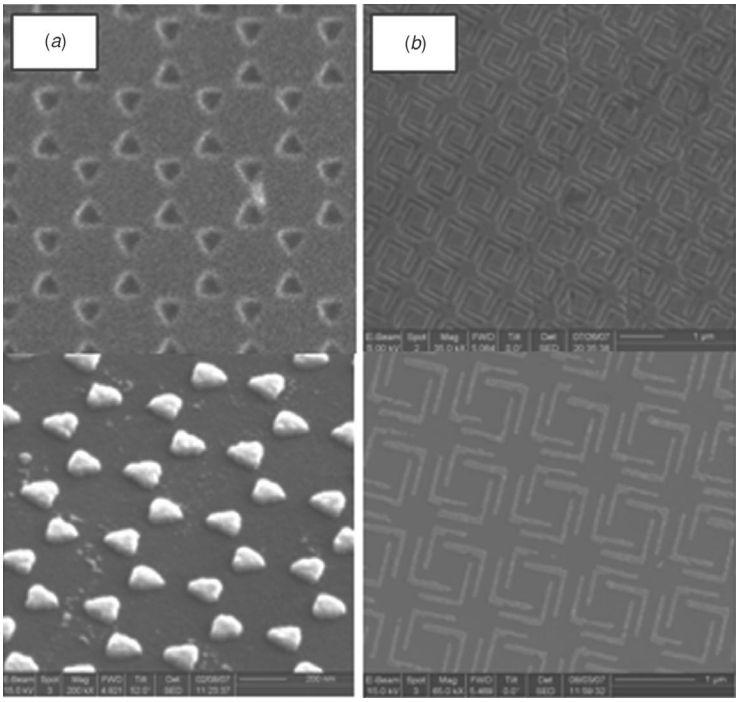


Figure 8.4. (*a*) A silver sulfide stamp with a pattern of 70-nm triangles made by FIB along with the results stamping shown below it. (*b*) A silver sulfide stamp made by embossing against a silicon master along with sub-100-nm stamping results below it.

The setup we have used to date is relatively simple and attests to the robust nature of the process. Shown in Figure 8.5, is the setup used for stamping. The stamp is mounted on a single-axis stage that is used to feed the stamp to the substrate. The stamp is attached to this assembly via an elastomer that provides 48 MPa of pressure at 20% compressive strain uniformly across the actual contact area between the stamp and the substrate to be imprinted. This substrate is mounted on a second stage that allows transverse or lateral positioning relative to the stamp. The nominal pressure (well below the yield stress of silver or copper sulfide) between the stamp and the substrate ensures a consistent contact between them during the progress of electrochemical imprinting. The electrical potential for electrochemical imprinting is controlled by a digital potentiostat (Gamry Instruments Model Reference 600) with a blocking electrode attached to the stamp as the cathode and the substrate being patterned as the anode. The process is performed in the chronoamperometry mode of the potentiostat, keeping the potential between anode and cathode constant, while monitoring the current flowing across them. To maintain the elevated temperature required for enhanced ionic conductivity in Cu_2S when performing electrochemical imprinting of copper, the stamp is enclosed in a circular heater set at 150 °C, while a second heater maintains the substrate at the same temperature.

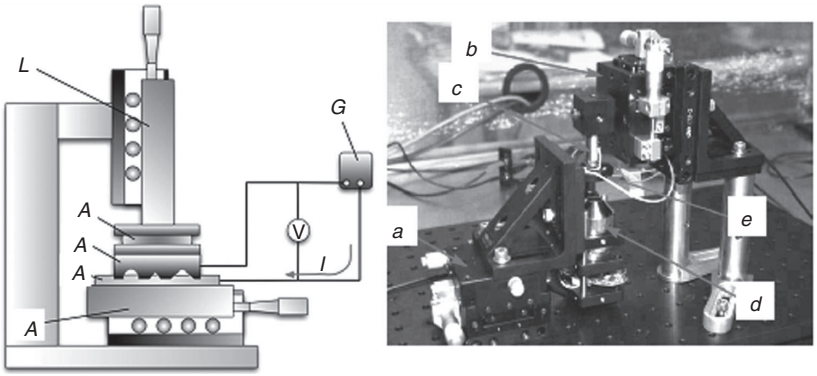


Figure 8.5. The S4 process setup. (*a*) X–Y stage with substrate holder, (*b*) dual stage for positioning and feeding stamp, (*c*) visco-elastic material and stamp holder, (*d*) silver sulfide stamp, (*e*) substrate with silver film, (*f*) potentiostat, (*g*) objective for observation. (Reproduced with permission from [21]. Copyright 2007 American Chemical Society.)

Our work has thus far concentrated on patterning of silver and copper films. These films are prepared by electron-beam evaporation of silver onto a 300-μm-thick glass cover slip or silicon wafer, cleaned using RCA1 solution. The films are deposited over a 10-nm Cr seed layer at a chamber pressure of 5×10^{-6} Torr and a stable rate of around 1 nm s^{-1}. Typical film thickness used by us ranges from about 50 to 500 nm, with 100-nm films being used for most of our experiments. For the copper films, prior to use, the substrates are cleaned using acetic acid (99.97^{+}) at room temperature for 1 min to remove any oxidation layers on the copper surface according to the technique proposed by Chavez et al. [34]. This step is necessary because the presence of a copper oxide layer was found to inhibit anodic dissolution at the interface and impede the S4 process.

8.4 PROCESS CAPABILITIES

An electrochemical imprinting process may be characterized by a number of parameters such as imprinting speed, resolution (finest feature transferred), stamping area, and stamp life. In addition, one may consider how these parameters change with process settings (e.g., dependence of imprinting speed on voltage or temperature, stamp life on imprinting force or voltage). Because this process is a relatively new, such dependencies of process output on operating conditions are still being studied. This discussion will therefore compare the relative performance of the process for the two materials, copper and silver. Further, we will not comment on stamping area because, to date, our stamping experiments are conducted on an improvized setup. The stamping area and, to a large extent, stamp life depend on the alignment and control of imprinting force between the stamp and the imprinted substrate.

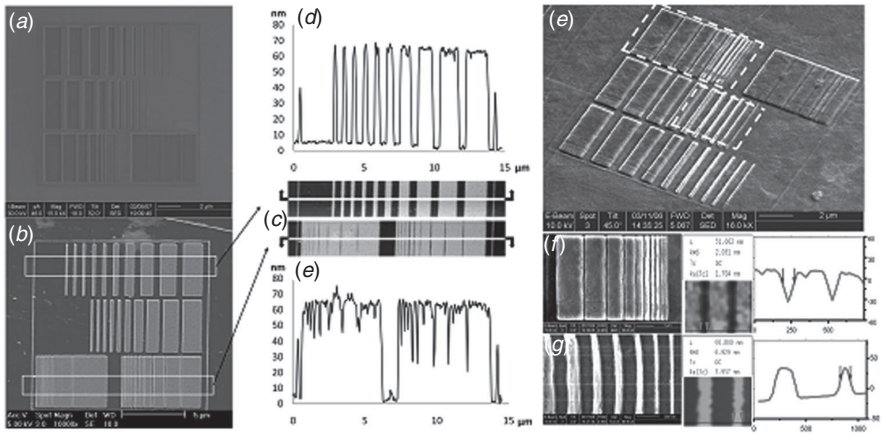


Figure 8.6. Line calibration results. (*a*), (*b*) SEM (scanning electron microscope) images of the Cu_2S stamp and resulting pattern imprinted in a 70-nm-thick Cu film. (*c*)–(*e*) AFM scan of pattern illustrating spacing and linewidths of 80 nm. (*f*), (*g*) SEM images of line patterns imprinted in a 100-nm thick Ag film. (*h*) AFM images and profiles illustrating spacing and linewidth of 50 nm. (Parts *a*–*e* are reproduced with permission from [22]. Copyright 2007 American Vacuum Society. Parts (*d*) and (*e*) are reproduced with permission from [21]. Copyright 2007 American Chemical Society.)

Q6 Q7 With both materials, silver and copper, we have been able to achieve sub-100-nm resolutions. Figure 8.6 shows bar-code patterns that allow us to measure both positive and negative features imprinted by the process. For silver, we observe good transfer of 50-nm patterns, while for copper, resolutions up to around 80 nm are observed. This also reflects our experience with the process. We have been able to routinely imprint sub-50-nm patterns into silver films, while imprinting of copper requires much greater care in the preparation of the stamp and during the process. This is due, in part, to the mechanical properties of the Cu_2S stamp. Cu_2S is brittle and hard. It tends to chip easily during the making or handling of the stamp. Not being malleable, it requires a higher contact force to elastically deform and make uniform electrical contact with the substrate to be patterned, making the finer features on stamp susceptible to damage.

With respect to stamping rates, imprinting silver with silver sulfide at room temperature is faster than imprinting copper at elevated temperatures. Figure 8.7 shows the imprinting speeds, i.e., speed of travel of the stamp into the film being patterned, for silver and copper. For imprinting of silver, imprinting speeds tend to be high (typically greater than $1\,nm\,s^{-1}$ and as high as $4\,nm\,s^{-1}$) and constant (see Figure 8.7). The average rates for copper are high enough to make the process economically competitive with other patterning process, but they are lower than those observed for silver. As shown in Figure 8.7 for patterning of 100-nm copper films at 150°C, we observed average of $0.295\,nm\,s^{-1}$ at 300 mV, $0.583\,nm\,s^{-1}$ at 600 mV, and $0.694\,nm\,s^{-1}$

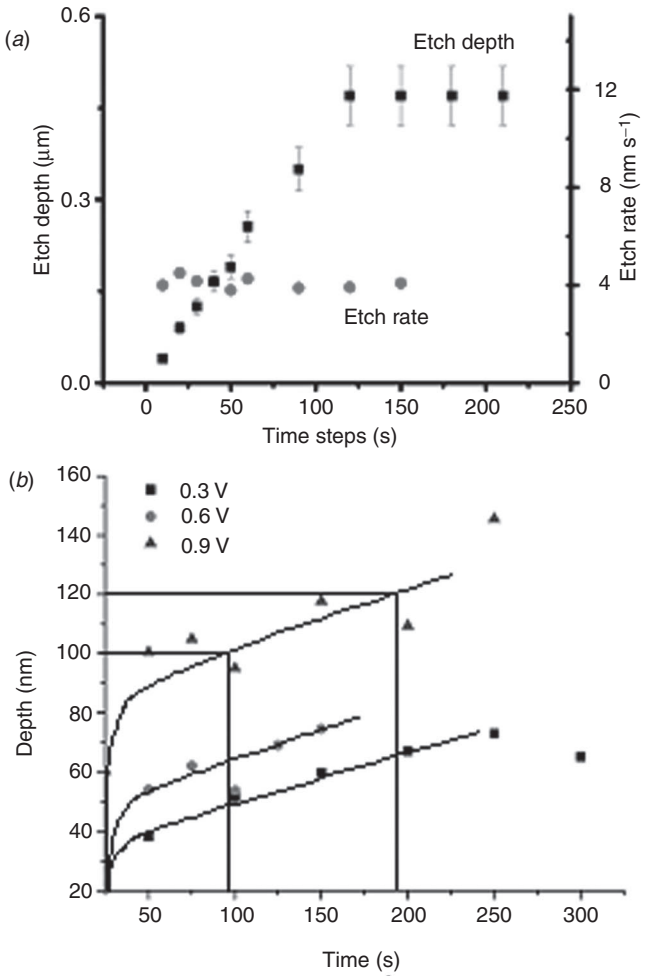


Figure 8.7. Travel of the stamp into the film as a function of time for (*a*) silver and (*b*) copper. Silver has a constant imprint speed or etch rate while copper has a high initial speed before settling to a constant lower imprint speed. (Part (*a*) is reproduced with permission from [21]. Copyright 2007 American Chemical Society. Part (*b*) is reproduced with permission from [22]. Copyright 2007 American Vacuum Society.)

at 900 mV. The etch rate is considerably higher at the onset, but appears to slow down significantly as the etching progresses to a constant value of about $0.2\,\mathrm{nm\,s^{-1}}$ as shown in the figure. The high etching rates at the start are attributable to the high copper concentration gradient and the thermal gradient at the contact interface at the onset of the process. During this time, the patterning rate is determined by the supply of mobile ions that, in turn, is determined by the rate of anodic dissolution.

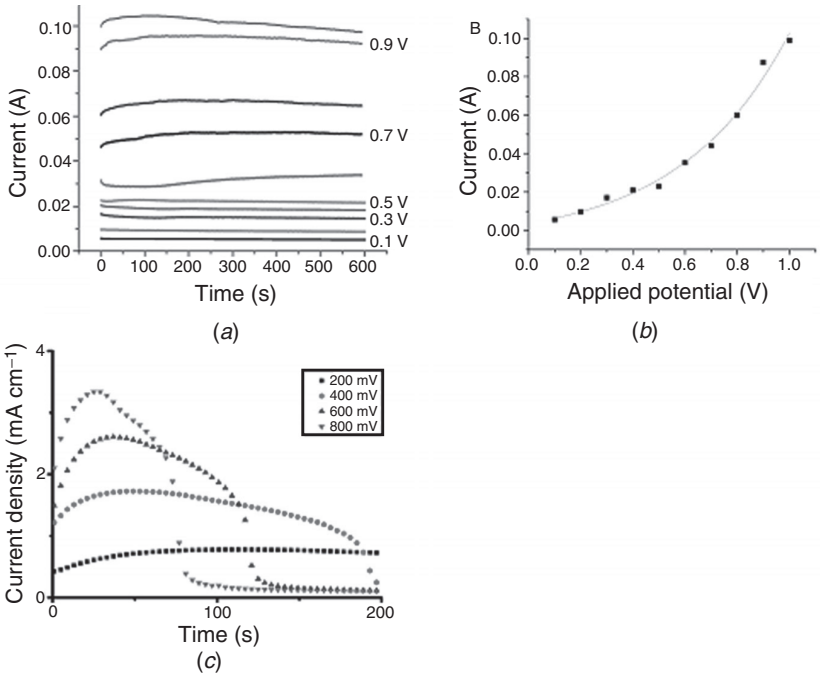


Figure 8.8. (*a*) The influence of increasing voltage bias on the current during imprinting of copper with copper sulfide. (*b*) Plot of initial (at 1 s) electronic current versus applied potential for the Cu_2S/Cu process. (Parts (*a*) and (*b*) are reproduced with permission from [22]. Copyright 2007 American Vacuum Society.) (*c*) Current density profiles while etching Ag with Ag_2S. In this case, the depletion of the film is clearly evident because of the high transference number for Ag_2S, indicating higher patterning speeds with increased voltages. (Part (*c*) is reproduced with permission from [21]. Copyright 2007 American Chemical Society.)

Subsequently, the imprinting rate is limited by the diffusion rate of the mobile Cu^{+2} ions in Cu_2S. This low etching rate compared to the Ag_2S/Ag system can, in part, be attributed to its low transference number at 150°C, 0.00003 as opposed to 0.11 in Ag_2S [35]. The dependence of imprinting speed on applied voltage is shown in Figure 8.8. In both cases, the increased current indicates an increased etch rate (though only a fraction of increase in current can be attributed to increased imprinting rates as the stamp materials are mixed, not pure, ionic conductors).

The S4 process is demonstrated to be repeatable from one imprint cycle to the next. Figure 8.9 shows the current profiles for imprinting of silver and copper. In both cases the current profile quickly settles to a steady state profile that displays three stages: a sharp ramp-up stage during which the anodic dissolution accelerates with activation of the electrochemical reaction at the anode, a second stage during which a slow decrease in current is observed as the stamp gets polarized, and a sharp decline when the supply of cations is depleted after the metallic film at the contact interface is completely etched away. Figure 8.10 shows a comparison of imprinting

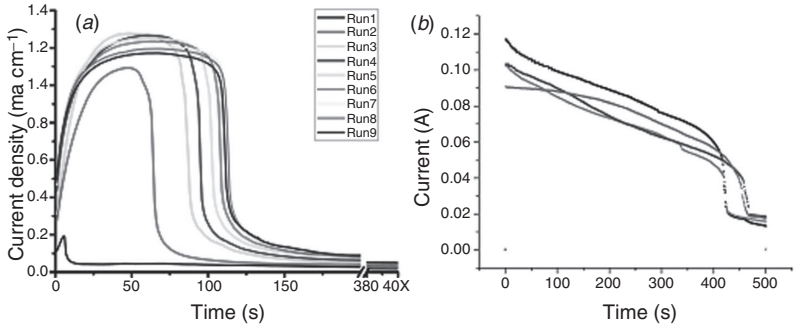


Figure 8.9. Current density and current profiles for repeated imprinting operations with the same stamp. (*a*) For Ag with Ag_2S, where after around two or three initial imprinting repetitions, the curve settles to a steady-state profile. (Reproduced with permission from [21]. Copyright 2007 American Chemical Society.) (*b*) For Cu_2S, where the current profile for etching 150-nm Cu films. (Reproduced with permission from [22]. Copyright 2007 American Vacuum Society.)

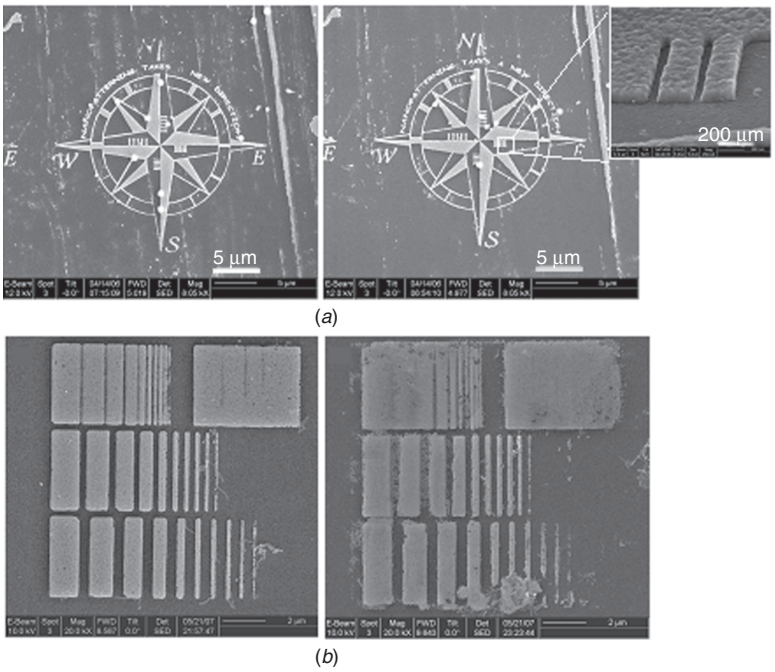


Figure 8.10. Imprints produced with multiple operations of the same stamp. (*a*) SEMs for imprinting Ag with an Ag_2S stamp show virtually no degradation of the imprint after nine repetitions. The two silver imprints correspond to the 1st and 9th curve of Figure 8.9*a*. (Reproduced with permission from [21]. Copyright 2007 American Chemical Society.) (*b*) SEMs of the 1st and 20th Cu imprint produced by the same Cu_2S stamp. Some of the smaller negative features vanish. The stamp was also damaged during the operation. (Reproduced with permission from [22]. Copyright 2007 American Vacuum Society.)

results as the imprinting stamp is used multiple times. With imprinting of silver, we have successfully used the same stamp tens of times, without seeing much degradation in stamp performance. Similarly, results were obtained with copper. However, because of working in a laboratory (not a clean room) environment with improvized equipment that does not allow us to fully control the process parameters such as imprinting force and to properly align the stamp with respect to the substrate, we see that some features on the stamp degrade and dust particles become embedded in the stamp. We expect that this will improve as we develop better imprinting equipment and stamp materials.

8.5 EXAMPLES OF ELECTROCHEMICALLY IMPRINTED NANOSTRUCTURES USING THE S4 PROCESS

We have successfully used the S4 process to make a number of structures. The barcodes of Figure 8.6 as well as the seals shown in Figure 8.11 have features in the sub-100-nm range and can be used for tagging and protection against counterfeiting. Figure 8.12 shows different metallic devices including electrodes for chemical sensors, such as meandering nanowires and interdigitated electrodes. Figure 8.13 shows experimental antennae construction for THz-frequency ranges. Figure 8.14 shows a silicon nitride substrate with a repetitive pattern of 70-nm silver triangles and the scheme for using it to create an SERS or LSPR substrate for chemical sensing.

Q3

Electrochemical imprinting of nanoscale patterns using S4 is relatively new technology, first reported in February of 2007 [21]. In spite of limited investment in the process technology, for example, stamp materials, or electrochemical imprinting equipment, resolutions in the range of 50–100 nm have been routinely achieved, suggesting good potential for industrial application. Our continuing work includes the exploration of new stamp materials that simultaneously give us favorable electrochemical and mechanical properties to facilitate efficient, large-area imprinting capabilities. Further, this exploration will allow us to enlarge the set of materials the S4 process is capable of addressing.

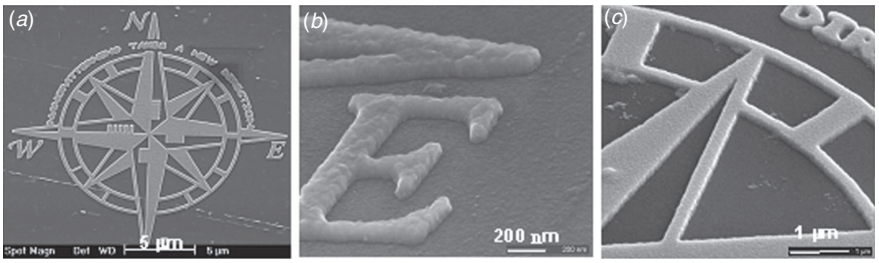


Figure 8.11. Examples of S4 stampings. (*a*)–(*c*) A high resolution silver stamping with feature definitions better than 50 nm can be used for nanoscale tagging of devices.

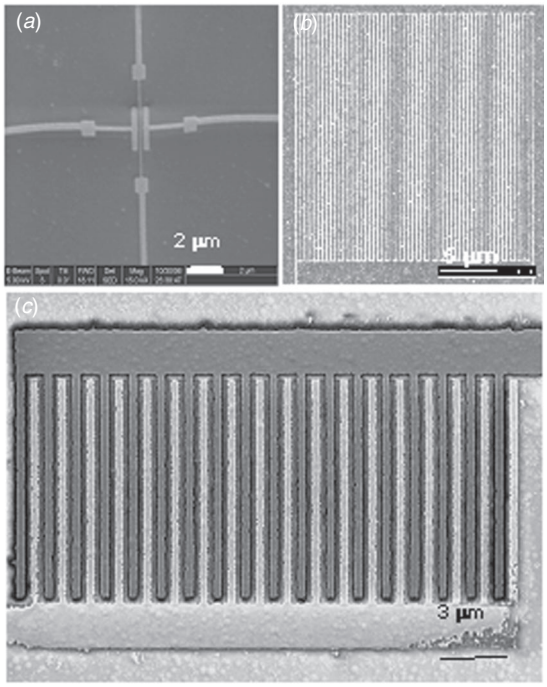


Figure 8.12. Different devices. (*a*) A 40-nm wire-based FET (field-effect transistor). (*b*) A meandering copper nanowire 70 nm wide and 1.35 mm long in a 20-micrometer square. (*c*) A set of interdigitated electrodes, 200 nm apart. The isolation between the two electrodes is evident by charging up of only the floating electrode in the SEM.

Fundamental to the process are the ionic transport phenomena at the contact interface and in the bulk of the stamp. The exploration of anodic dissolution at solid–solid Q8
interfaces has received relatively little attention (in comparison to fluid–solid) inter- Q9
faces. This is an area we are currently exploring to gain a better understanding of the process at the interface influence, the process dynamics, and the stamp surface (and hence, overall stamp life). Understanding the bulk transport of ions within the stamp plays an important part in understanding the process and use of the stamp. We have been developing models for ion transport based on the Nernst–Planck equation to predict (cat)ion concentration in the stamp at the end of each imprinting cycle. This can then be used to gauge the polarization of the stamp material and develop an idea of how often and for how long the stamp needs to be depolarized by using a blocking electrode as the anode.

Research is also continuing into experimental studies of the role of imprinting pressure, voltage, and temperature on the speed of imprinting process, surface quality and fidelity of the imprinted structures, and life of the stamp. In addition, we are developing an instrumented, precision imprinting press for conducting such process characterization experiments.

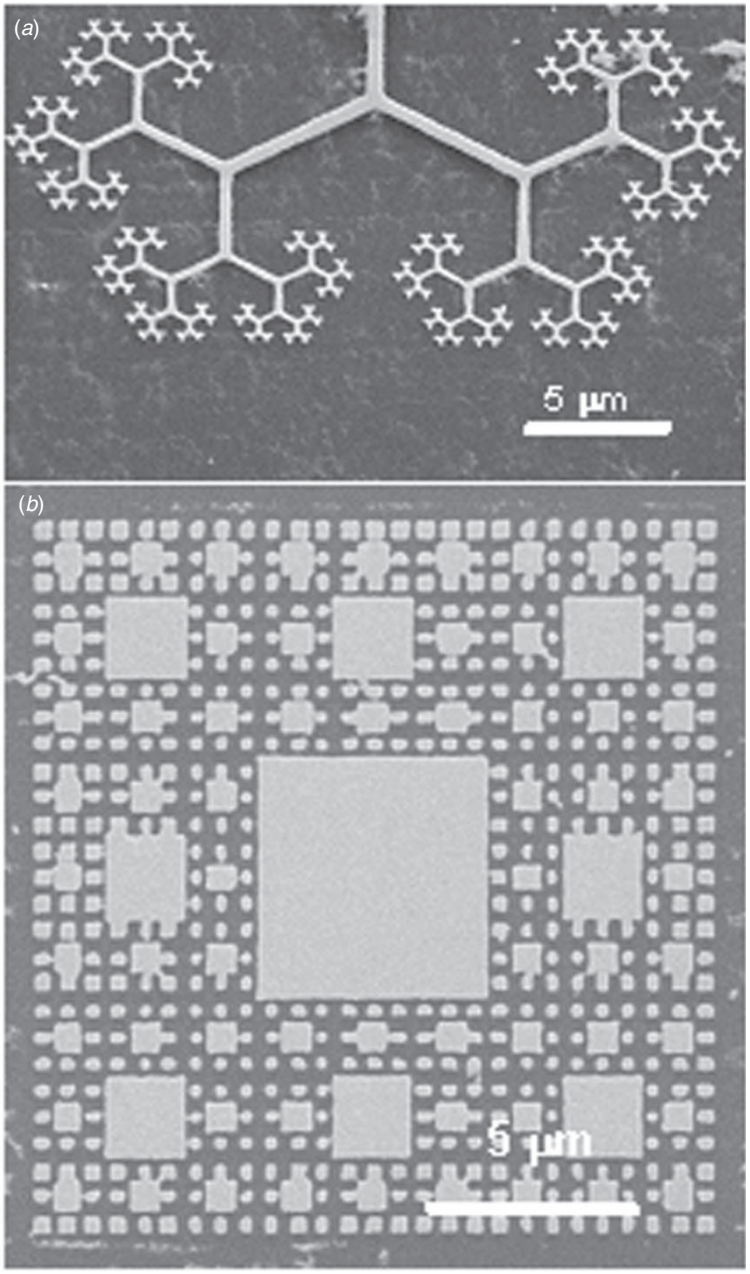


Figure 8.13. (*a*) A Pythagoras tree antenna with 70-nm features. (*b*) A 2D Menger sponge antenna for frequencies up to 40 THz.

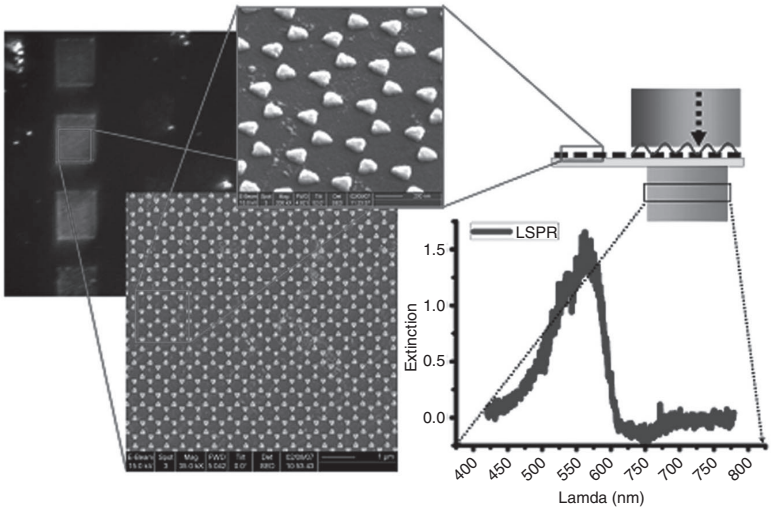


Figure 8.14. Large areas of repetitive patterns made by the S4 process can be used for SERS or LSPR-based sensing of chemicals. The middle shows silver triangles of silver patterned on glass. The image to the right shows dark-field images of patterned areas illuminated with white light. The figure to the right schematically shows how such substrates excite surface plasmons that extinguish certain frequencies from the transmitted spectrum.

ACKNOWLEDGMENTS

This research was supported by NSF through the Center for Chemical-Electrical-Mechanical Manufacturing Systems (Nano-CEMMS) under Grant DMI-0312862, the Office of Naval Research under grant N00173-07-G013, and the University of Illinois through the Grainger Foundation grant. We are grateful that part of this work was carried out in the Center for Microanalysis of Materials, University of Illinois, which is partially supported by the US Department of Energy under grant DEFG02-ER45439.

REFERENCES

1. Madou, M. (2002) *Fundamentals of Microfabrication*, 2nd edn., CRC Press, New York.
2. Rickerby, J. and Steinke, J. (2002) Current trends in patterning with copper. *Chem. Rev.* **102**, 1525–1549.
3. Felmet, K., Loo, Y., and Sun, Y. (2004) Patterning conductive copper by nanotransfer printing. *Appl. Phys. Lett.* **85**, 3316–3318.
4. Ruska, W. S. (1987) *Microelectronic Processing*, McGraw-Hill, New York.
5. Ebbesen, T. W., Lezec, H. J., Ghaemi, H. F., Thio, T., and Wolff, P. A. (1998) Extraordinary optical transmission through sub-wavelength hole arrays. *Nature* **391**, 667–669.

Q4

6. Maiera, S. A and Atwater, H. A. (2005) Plasmonics: localization and guiding of electromagnetic energy in metal/dielectric structures. *J. Appl. Phys.* **98**, 011101.
7. Andricacos, P. C., Uzoh, C., Dukovic, J. O., Horkans, J., and Deligianni, H. (1998) Damascene copper electroplating for chip interconnections. *IBM J. Res. Dev.* **42**, 567–574.
8. Schmid, G. M., Stewart, M. D., Wetzel, J., Palmieri, F., Hao, J., Nishimura, Y., Jen, K., Kim, E. K., et al. (2006) Implementation of an imprint damascene process for interconnect fabrication. *J. Vac. Sci. Technol. B* **24**, 1283.
9. Tamirisa, P. A., Levitin, G., Kulkarni, N. S., and Hess, D. W. (2007) Plasma etching of copper films at low temperature. *Microelectron. Eng.* **84**, 105–108.
10. Steinhögl, W., Schindler, G., Steinlesberger, G., Traving, M., and Engelhardt, M. (2005) Comprehensive study of the resistivity of copper wires with lateral dimensions of 100 nm and smaller. *J. Appl. Phys.* **97**, 023706.
11. Steinbruchel, C. (1995) Patterning of copper for multilevel metallization: reactive ion etching and chemical-mechanical polishing. *Appl. Surf. Sci.* **91**, 139–146.
12. Schwartz, G. C. and Schaible, P. M. (1983) Reactive ion etching of copper films. *J. Electrochem. Soc.* **130**, 1777.
13. Ohno, K., Sato, M., and Arita, Y. (1996) Reactive ion etching of copper films in a $SiCl_4$, N_2, Cl_2, and NH_3 mixture. *J. Electrochem. Soc.* **143**, 4089.
14. Howard, B. J. and Steinbruchel, C. (1991) Reactive ion etching of copper in SCl_4-based plasmas. *Appl. Phys. Lett.* **59**, 19.
15. Howard, B. J. and Steinbruchel, C. (1994) Reactive ion etching of copper with BCl_3 and $SiCl_4$: plasma diagnostics and patterning. *J. Vac. Sci. Technol. A* **12**, 1259–1264.
16. Lee, S. and Kuo, Y. (2001) Chlorine plasma/copper reaction in a new copper dry etching process. *J. Electrochem. Soc.* **148**, G524–G529.
17. Lim, J., Kim, N., and Chang, E. (2004) Electrochemical patterning of copper using microcontact printing. *J. Electrochem. Soc.* **151**, C455–C458.
18. Hulteen, J. C. and Van Duyne, R. P. (1995) Nanosphere lithography: a materials general fabrication process for periodic particle array surfaces. *J. Vac. Sci. Technol. A* **13**, 1553.
19. Trimmer, A. L., Hudson, J. L., Kock, M., and Schuster, R. (2003) Single-step electrochemical machining of complex nanostructures with ultrashort voltage pulses. *Appl. Phys. Lett.* **82**, 3327–3329.
20. Bhattacharyya, B., Doloi, B., and Sridhar, P. S. (2001) Electrochemical micro-machining: new possibilities for micro-manufacturing. *J. Mater. Process. Technol.* **113**, 301–305.
21. Hsu, K. H., Schultz, P. L., Ferreira, P. M., and Fang, N. X. (2007) Electrochemical nanoimprinting with solid-state superionic stamps. *Nano Lett.* **7** (2), 446–451.
22. Schultz, P. L., Hsu, K. H., Fang, N. X., and Ferreira, P. M. (2007) Solid-state electrochemical nanoimprinting of copper. *J. Vac. Sci. Technol. B* **25 (6)**, 2419–2424.
23. Chou, S. Y., Krauss, P. R., and Renstrom, P. J. (1996) Nanoimprint lithography. *J. Vac. Sci. Technol. B* **14**, 4129–4133.
24. Chou, S. Y., Krauss, P. R., Zhang, W., Guo, L., and Zhuang, L. (1997) Sub-10 nm imprint lithography and applications. *J. Vac. Sci. Technol. B* **15**, 2897–2904.
25. Terabe, K., Nakayama, T., Hasegawa, T., and Aono, M. (2002) Ionic/electronic mixed conductor tip of a scanning tunneling microscope as a metal atom source for nanostructuring. *Appl. Phys. Lett.* **80**, 4009–4011.

26. Terabe, K., Nakayama, T., Hasegawa, T., and Aono, M. (2002) Formation and disappearance of a nanoscale silver cluster realized by solid electrochemical reaction. *J. Appl. Phys.* **91**, 10110–10114.
27. Lee, M., O'Hayre, R., Prinze, F. B., and Gur, T. M. (2004) Electrochemical nanopatterning of Ag on solid-state ionic conductor $RbAg_4I_5$ using atomic force microscopy. *Appl. Phys. Lett.* **85**, 3552–3554.
28. Agrawal, R. and Gupta, R. (1999) Superionic solids: composite electrolyte phase—an overview. *J. Mater. Sci.* **34**, 1131–1162.
29. Miyatani, S. (1956) Point contact of Pt and γ-Cu2S. *J. Phys. Soc. Japan* **11**, 1059–1063.
30. Allen, L. and Buhks, E. (1984) Copper electromigration in polycrystalline copper sulfide. *J. Appl. Phys.* **56**, 327–335.
31. Hirahara, E. (1951) The physical properties of cuprous sulfides-semiconductors. *J. Phys. Soc. Japan* **6**, 422–427.
32. Hebb, M. H. (1952) Electrical conductivity of silver sulfide. *J. Chem. Phys.* **20** (1), 185–190.
33. Wagner, C. (1953) Investigations on silver sulfide. *J. Chem. Phys.* **21**, 1819–1827.
34. Chavez, K. and Hess, D. (2001) A novel method of etching copper oxide using acetic acid. *J. Electrochem. Soc.* **148**, 640–643.
35. Pauporte, T. and Vedel, J. (1998) Electrical properties of a non-stoichiometric copper sulfide/solid electrolyte interface. *Solid State Ion.* **109**, 125–134.

LIST OF QUERIES

Q1. Please provide complete first names of all the authors.

Q2. Please check whether the expanded forms of ”STM” and ”AFM” are OK as supplied here.

Q3. Please expand ”SERS” and ”LSPR” here.

Q4. Please check the supplied article title and page range in reference [5].

Q5. Please expand ”PR” here.

Q6. Please check whether the expanded form of SEM is OK as supplied.

Q7. The credit line for parts (*d*) and (*e*) in figure caption 8.6 is appearing twice, first for [22] and then for [21]. Could you please check it for correctness? Please also check the artwork and confirm that the figure parts are labeled correctly.

Q8. Please check whether the expanded form of FET is OK as supplied.

Q9. Please check the artwork for Figure 8.12 and confirm that the figure parts are labeled correctly.

Electrostatically Actuated Cantilever With SOI-MEMS Parallel Kinematic XY Stage

Jingyan Dong, *Member, ASME*, and Placid M. Ferreira, *Member, ASME*

***Abstract*—This paper presents the design, analysis, fabrication, and characterization of an active cantilever device integrated with a high-bandwidth 2-DOF translational (XY) micropositioning stage. The cantilever is actuated electrostatically through a separate electrode that is fabricated underneath the cantilever. Torsion bars that connect the cantilever to the rest of the structure provide the required compliance for the cantilever's out-of-plane rotation. The active cantilever is carried by a micropositioning stage, which enables high-bandwidth scanning to allow manipulation in three dimensions. The design of the microelectromechanical system stage is based on a parallel kinematic mechanism (PKM). The PKM design decouples the motion in the X- and Y-directions while allowing for an increased motion range with linear kinematics in the operating region (or workspace). The trusslike structure of the PKM also results in increased stiffness and reduced mass of the stage. The integrated cantilever device is fabricated on a silicon-on-insulator (SOI) wafer using surface micromachining and deep reactive ion etching processes. The actuation electrode of the cantilever is fabricated on the handle layer, while the cantilever and the XY stage are at the device layer of the SOI wafer. Two sets of electrostatic linear comb drives are used to actuate the stage mechanism in the X- and Y-directions. The cantilever provides an out-of-plane motion of 7 μm at 4.5 V, while the XY stage provides a motion range of 24 μm in each direction at the driving voltage of 180 V. The resonant frequency of the XY stage under ambient conditions is 2090 Hz. A high quality factor ($\sim$210) is achieved from this parallel kinematic XY stage. The fabricated stages will be adapted as chip-scale manufacturing and metrology devices for nanomanufacturing and nanometrology applications. [2008-0278]**

***Index Terms*—Active cantilever, comb drive, micropositioning stage, parallel kinematic mechanism (PKM), tilt-plate actuator.**

I. Introduction

THE development of microcantilever-based devices has played a key role in advances of nanotechnology over the last two decades. By using a microcantilever with a sharp tip as a sensor, these devices can sense extremely small physical signals, such as tip-sample tunneling current in scanning tunneling microscope [1], [2] and interatomic-force atomic force microscope [3]. Their capability of manipulation at the atomic scale, together with the capability of sensing a variety of physical signals in diverse environments, brings a dramatic impact in the fields of biology, materials science, tribology, surface physics, and medical diagnosis [4], [5]. By vibrating the cantilever and detecting the change of its resonant frequency and vibrational magnitude, cantilever-based devices are used as chemical sensors [6], [7] to detect some specific chemicals absorbed by the pretreated cantilever. Microcantilever-based devices are also widely used in micro-/nanofabrication applications, such as dip pen lithography [8], thermal embossing [9], [10], local oxidation, and resist exposure [11].

Although in many applications cantilevers are used as passive sensors that are bent by an external force, active cantilevers offer additional advantages and capabilities, such as self-excitation and individual tip–substrate separation adjustment, particularly when used in a cantilever array. Different technologies have been used to provide actuation to microcantilever devices, including electrothermal actuators [12], electromagnetic actuators [13], piezoelectric actuators [14], [15], shape memory alloy actuators [16], and electrostatic actuators [17]. Among these actuation technologies, electrostatic actuators offer some unique features, compared with other actuation techniques, because of their simplicity and the ease with which their fabrication is integrated with that of the rest of the structure. Unlike the other actuation technologies, electrostatic actuators avoid extra processing steps and additional materials, such as shape memory alloys, piezoelectric films/actuators, electrically heated resistors, or magnets/coils.

A typical cantilever-based instrument or system is configured with a microscale cantilever as the sensor or the functional manipulator, and a mesoscale flexure-based piezoelectric-actuator-driven nanopositioning stage. Although the cantilever can work at extremely high frequency ($>$ 10 kHz), the relatively slow nanopositioner limits the overall scanning speed and becomes the bottleneck in a high-throughput system due to its excess mass. Furthermore, in such a configuration, when using multiple cantilevers, all the cantilevers undergo the same relative displacement with respect to the substrate that they are processing, making it difficult to exploit the inherent parallelism of the multicantilever system. In contrast, a microscale microelectromechanical systems (MEMS) positioning device can achieve a much higher bandwidth due to scaling effects. The widely used XY stage designs [18]–[22] include four identical comb-drive structures that are placed around the end effector, with each of them being perpendicular to its neighbor. The end effector or moving platform is connected to the four comb actuators by long slender beams. When the stage is actuated in the

Manuscript received November 11, 2008; revised December 23, 2008 and February 8, 2009. First published May 8, 2009; current version published June 3, 2009. This work was supported by the National Science Foundation through the Center for Nanoscale Chemical Electrical and Mechanical Manufacturing Systems under Award DMI 0328162 and through Grant Awards DMI 0422687, and CMMI 0800863. Subject Editor D. Elata.

J. Dong is with the Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695-7906 USA (e-mail: jdong@ncsu.edu).

P. M. Ferreira is with the Department of Mechanical Science and Engineering, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA (e-mail: pferreir@uiuc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JMEMS.2009.2020371

X-direction, the long beam along the Y-direction acts as a leaf spring to accommodate the motion of the X-axis and vice versa. Due to such a coupled structure design, the crosstalk between the X- and Y-axes can cause problems such as side instability of comb drives and limited motion range. Reducing the stiffness of the leaf springs reduces the crosstalk, alleviating some of its negative effects. However, the decreases in stiffness of the beams lead to nondeterministic motions along with undesirable end-effector rotations in the XY plane. Additionally reduced stiffness leads to lower resonant frequencies and complex dynamics with multiple modes [23] in a fairly narrow frequency band. Other than the aforementioned designs, serial kinematic designs are introduced by Takahashi *et al.* [24], [25]. The 2 DOF is realized by the serial combination of two single DOF systems, in which the inner axis is embedded into the moving frame of the outer axis. Thus, the moving inertia load of the two axes is different. Also, the additional mass significantly decreases the natural frequency and response time of the outer axis. Additionally, electrical isolation can be a problem. Monolithic parallel kinematic mechanisms (PKMs) are well suited for silicon-based micropositioning devices. Compared with serial kinematic designs, parallel kinematic mechanisms can significantly increase the natural frequency of the designed system due to their high structure stiffness and low inertia, which accrue from their trusslike structures. Furthermore, if appropriately designed, PKMs can result in configurations with nearly complete decoupling of the actuation effect and better position accuracy. Micropositioning stages based on PKM with different degrees of freedom [26]–[31] demonstrate good performance capabilities in their motion range, bandwidth, and resolution [26]–[32].

In this paper, an electrostatically actuated active cantilever device that enables operation in 3 DOF is designed and fabricated. The chip-scale cantilever device is expected to address applications such as high-throughput nanoscale metrology, imaging, and manufacturing. In this device design, a parallel kinematic micropositioning stage is used to provide the cantilever with high-bandwidth motion in the XY plane. Linear comb drives, along with folded springs, provide the required linear actuation. Parallelogram four-bar linkages absorb the coupling motion between the two axes while simultaneously confining the orientation of the end effector. The cantilever is connected to the stage through torsion bars that provide the required compliance for the cantilever's out-of-plane motion. The actuation is provided by a tilt-plate actuator that is on the top of the cantilever. The overall device is fabricated on a silicon-on-isolator (SOI) die with a 50-μm-thick device layer and a 3-μm-thick buried oxide (BOX) layer. The high-aspect-ratio structures, such as comb fingers, torsion bars, and cantilevers, are fabricated by deep reactive ion etching (DRIE). The actuation electrode for the out-of-plane motion of the cantilever is realized on the handle layer, while the cantilever itself and the XY stage are at the device layer of the SOI die. The handle layer beneath the device is etched away, except for the tilt-plate electrode for actuating the cantilever. This enables the stage to be used in applications that require access to the positioning platform from both the top and bottom directions. The fabricated device provides an out-of-plane cantilever motion of

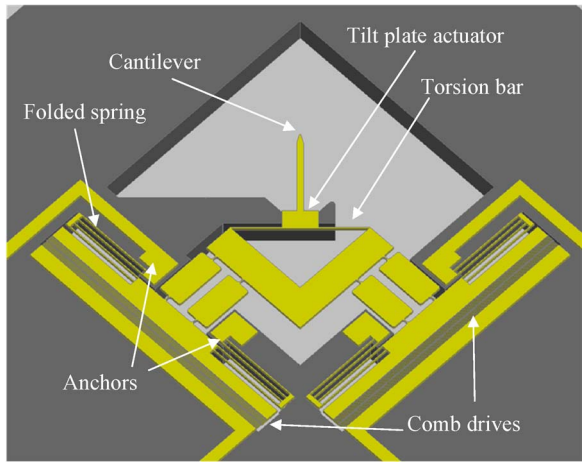


Fig. 1. Schematic diagram of the active cantilever device with parallel kinematic XY stage.

7 μm at 4.5 V, with a motion range of more than 24 μm in the X- and Y-directions at a driving voltage of 180 V. The resonant frequency of the XY scanning under ambient conditions is 2090 Hz. A high quality factor ($\sim$210) is achieved from this XY positioning stage.

II. Device Design and Analysis

Fig. 1 shows the design of the active cantilever device with parallel kinematic high-bandwidth MEMS XY positioner. In this design, the cantilever is connected to a parallel kinematic micropositioning stage through two torsion bars. The torsion bars provide the rotary compliance of the cantilever structure that enables the out-of-plane motion of the cantilever. A rectangular plate that is placed at the root of the cantilever acts as one electrode of a tilt-plate actuator. The counter electrode of this tilt-plate actuator is created in the handle layer that is located underneath the cantilever plate. The separation gap between two plates is defined by the thickness of the sacrificial layer (BOX layer) of SOI wafers. When a potential difference is applied to the cantilever plate (ground) and the handle-layer plate (voltage), the electrostatic force from the tilt-plate actuator generates a torque for rotating the torsion bars, thus displacing the cantilever in the Z-direction. Since the tilt-plate actuator is located very close to the torsion bars, the lever-arm effect produces a relatively large displacement at the cantilever tip, in spite of the small separation gap between two plates.

The cantilever is carried by a parallel kinematic micropositioning XY stage. The design of the XY stage is schematically shown in Fig. 2. In this design, there are two independent kinematic chains that connect the end effector to the base. Each kinematic chain includes two serially connected components: a prismatic joint that provides pure translational motion and a parallelogram four-bar-linkage mechanism that provides rotary displacement while holding the orientation of the end-effector invariant. These two chains are placed perpendicular to each other so as to kinematically decouple the two actuated joints to the maximum extent possible. The two kinematic chains are identical to each other, which results in identical dynamics of the stage along any direction in the XY plane. When the stage

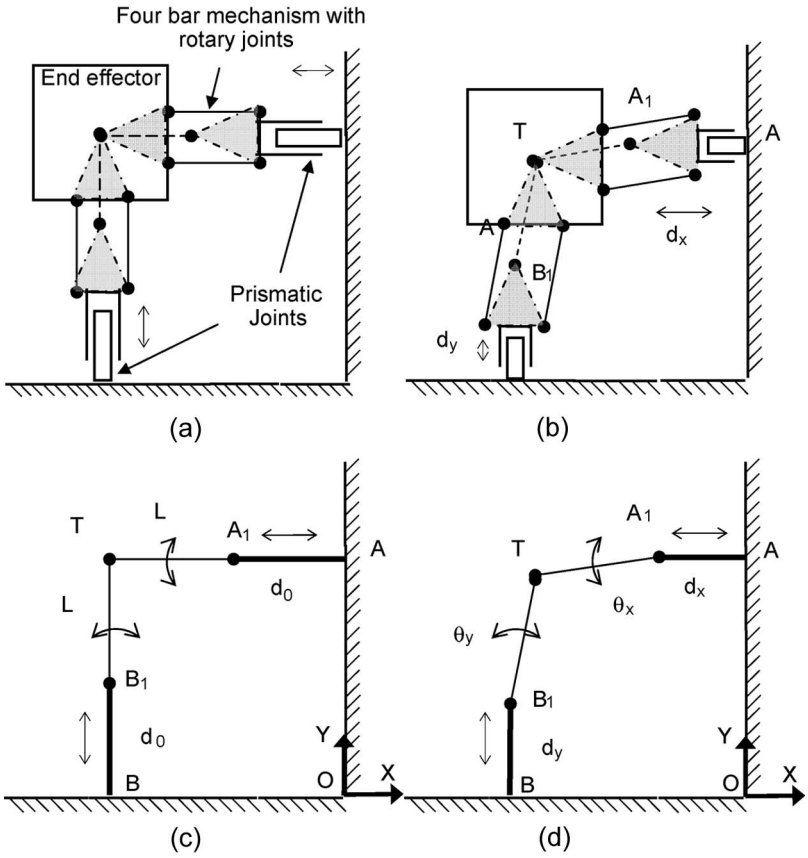


Fig. 2. (a) Addition of links between the base and connector of the parallelogram four-bar linkage does not change the mobility of the system, provided that the links are parallel to the crank (and follower) and of the same length. (b) System configuration under displacements at two prismatic joints. (c), (d) Equivalent linkage of the parallel kinematic XY stage for the purpose of analysis at the nominal position and under actuation.

is actuated in one direction by the prismatic joint of that chain, the resulting motion of the end effector is accommodated by the other kinematic chain by an angular displacement at the parallelogram four-bar mechanism.

To simplify the analysis of the system, we use a virtual link. Adding a link between the base and connector of a parallelogram four-bar linkage does not change its mobility, provided that this link has the same length as the crank and is parallel to it. This is shown in Fig. 2. Links are added to the two four-bar linkages so that they meet at the same point, i.e., T on the end effector. The simplified kinematic model is shown in Fig. 2(c) and (d). Links AA_1 and BB_1 represent two prismatic joints with an initial length d_0, and this length will be changed along with the actuation effect. Links TA_1 and TB_1 represent the virtual links with a fixed length L. Here, the motion of point T completely represents the motion of the end effector because it only undergoes pure translation.

It is now easy to find the relationship between actuation displacement, which is linear displacement of two prismatic joints, and the displacement of the stage. Referring to Fig. 2, if the coordinates of point O are chosen as the origin (0, 0), then the position where the kinematic chains are connected to the base is $A(0, d_0 + L)$, $B(-d_0 - L, 0)$. The nominal position for the table in this coordinate system is $(-d_0 - L, d_0 + L)$. After the prismatic joints are actuated to d_x and d_y, the coordinates at the end of the joints are $A_1(-d_x, d_0 + L)$, $B_1(-d_0 - L, d_y)$. The coordinates of the new position of table T can now be solved as the length of the second joint remains the same. By letting the position of the stage be (x, y), then $|TB_1| = |TA_1| = L$, which satisfies the relationships given in

$$\begin{aligned}(x + d_x)^2 + (y - d_0 - L)^2 &= L^2\\ (x + d_0 + L)^2 + (y - d_y)^2 &= L^2 \qquad (1)\\ d_x &= \sqrt{L^2 - (y - d_0 - L)^2} - x\\ d_y &= y - \sqrt{L^2 - (x + d_0 + L)^2}. \qquad (2)\end{aligned}$$

The angular displacement of the second link that is a four-bar linkage is

$$\begin{aligned}\theta_x &= \sin^{-1}\frac{\Delta y}{L} = \sin^{-1}\frac{d_0 + L - y}{L}\\ \Delta\theta_y &= \sin^{-1}\frac{\Delta x}{L} = \sin^{-1}\frac{x + d_0 + L}{L} \qquad (3)\end{aligned}$$

where Δx and Δy are the displacements of the end effector in the X- and Y-directions, respectively. Differentiating (2) with respect to x, y at operation points $T_0(-d_0 - L, d_0 + L)$, we obtain

$$\begin{aligned}\begin{bmatrix}\Delta d_x\\ \Delta d_y\end{bmatrix} &= \begin{bmatrix}-1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix}\\ \begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix} &= J\begin{bmatrix}\Delta d_x\\ \Delta d_y\end{bmatrix} = \begin{bmatrix}-1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}\Delta d_x\\ \Delta d_y\end{bmatrix}. \qquad (4)\end{aligned}$$

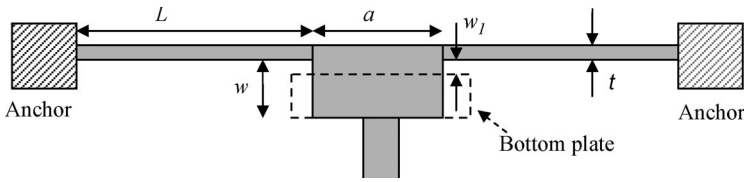


Fig. 3. Torsion bars and tilt plate for the cantilever.

The diagonal Jacobian matrix in (4) suggests a decoupled motion in the X- and Y-directions around the nominal point, when the device undergoes a small displacement relative to the overall dimensions of the stage (which is just the case for this MEMS system). One actuator will generate displacement in the X-direction and the other in the Y-direction. The effect of crosstalk between different axes (at the first order of approximation) is zero. As shown in Fig. 1, for the XY stage, the rotary joints around the four-bar mechanism are implemented by flexure hinges, and the prismatic joint is realized by the linear comb drive and the folded-spring suspension structure.

III. Structural Design and Analysis

We first analyze the out-of-plane motion of the cantilever. As previously mentioned, the z motion is obtained by the rotation of the cantilever, which is permitted by two torsion bars that connect the cantilever to the end effector of the XY stage. Fig. 3 shows a schematic of torsion bars and parallel-plate actuator for the cantilever. The rotational stiffness of the torsion bars is given by [33]

$$K = \frac{T_k}{\varphi} = \frac{2\beta h t^3 G}{L} \tag{5}$$

where T_k is the torque, φ is the angular displacement, G is the shear modulus of the material of torsion bars, L is the length of a torsion bar, h is the height, t is the width of the torsion bar, and β is a numerical factor depending on the ratio of h/t. For the torsion bars used in our design, we have beam width $t = 5\ \mu$m, $L = 800\ \mu$m, $h = 50\ \mu$m, and $\beta = 0.291$. The shear modulus of single-crystal silicon $\langle 100 \rangle$, as has been used for device fabrication, is about 51 GPa. Therefore, the rotational stiffness of our torsion bar is 2.3×10^{-7} N $\cdot$ m/rad.

The actuation torque at a certain rotation angle can be derived by integrating the actuation effect over the overlapped area of the tilt plate. The torque generated from the electrostatic force is

$$\begin{aligned} T_a &= \int_{w_1}^{w} \frac{\varepsilon V^2 a x dx}{2(g-\varphi x)^2} \\ &= \frac{\varepsilon V^2 a}{2\varphi^2}\left(\ln\frac{g-\varphi w}{g-\varphi w_1} + \frac{\varphi g(w-w_1)}{(g-\varphi w)(g-\varphi w_1)}\right) \end{aligned} \tag{6}$$

where ε is the electrical permittivity, V is the actuation voltage, g is the gap between two plates, a is the width overlap of the two plates, and w_1 and w are the start and the end of the overall lap from the rotational center. The rational angle of the tilt plate at a certain voltage can be derived through torque balance equation $T_a = T_k = K\varphi$ and solved numerically. In our design, we have $a = 550\ \mu$m, $w = 300\ \mu$m, and $w_1 = 100\ \mu$m. We can get a maximum angular defection of 4.35×10^{-3} rad at about 5.5 V. The corresponding maximum displacement at the center of the electrode is 0.87 μm. The lever-arm effect will magnify this displacement at the tip of the cantilever and result in about 7-μm out-of-plane displacement.

Next, we consider the in-plane motion of the stage. One-dimensional circular flexure hinges are used in the stage structure as rotary joints. The rotary stiffness of the hinge is given by [34]

$$K_z = \frac{T_k}{\varphi} = \frac{2Eh}{9\pi}\sqrt{\frac{t^5}{R}} \tag{7}$$

where E is the Young's modulus of the material of a flexure hinge, t is the neck thickness of the flexure hinge (6 μm), R is the radius of a flexure hinge at the neck ($= 300\ \mu$m), and h is the height of the flexure hinge or the device thickness ($= 50\ \mu$m). The maximum bending torque that can be applied to a flexure hinge and, consequently, the maximum rotary deflection are given by [26]. The Young's modulus of the silicon-based 1-D circular flexure hinge used in the stage mechanism is 131 GPa, and its elastic limit is about 7000 MPa. Hence, the rotational stiffness and the maximum angular deflections of the hinges can be given by

$$K_z = 2.36 \times 10^{-6}\ \text{N} \cdot \text{m} \cdot \text{rad}^{-1} \tag{8}$$

$$\alpha_{\max} = 0.68\ \text{rad}. \tag{9}$$

The length of the four-bar structure is 1 mm, which indicates a maximum 630-μm displacement of the four-bar-linkage mechanical structure. Factors such as suspension structure and limited actuating forces and stroke prevent us from reaching this limit.

The displacement of a linear comb drive is defined by the stiffness of the folded spring as the suspension structure, as well as the actuation force that it can provide. In our design, a folded spring is used to support the rotor and the table. The designed folded spring has a large compliance in the actuation direction for large displacements and a much higher stiffness in the lateral direction so as to prevent side instabilities. From the beam deflection theory [35], the stiffness values of a clamped–clamped beam in the motion and lateral directions, i.e., k_d and k_l, respectively, can be expressed as follows:

$$k_d = 2Eht^3/L^3 \quad k_l = 2Eht/L \tag{10}$$

where h is the height of the beam or the device thickness ($= 50\ \mu$m), t is the width of the beam (15 μm), and L is

the length of the beam ($= 1.375$ mm). In our design, two clamped–clamped beams are used in series to form a folded spring. The first beam connects the anchors to an intermediate truss, and the second one connects the truss to the rotor. The lengths of the two beams are the same to prevent an undesirable parasitic motion. We obtain stiffness values in the displacement direction as $k_d = 17.0\ \mathrm{N \cdot m^{-1}}$ and in the lateral direction as $k_l = 142\,909\ \mathrm{N \cdot m^{-1}}$, resulting in the stiffness ratio of $k_l/k_d = 8403$.

The linear-comb-drive actuator provides force to overcome the stiffness from the folded spring and flexure hinges under an actuation voltage V. The electrostatic force is given by

$$F = n\frac{\varepsilon_0 h V^2}{g} \tag{11}$$

where n is the number of finger pairs ($= 191$), h is the height of a finger (50 μm), g is the gap between two neighboring fingers (5 μm), and ε_0 is the electrical permittivity. Thus, the linear comb drive can generate a force of 380 μN at 150 V.

The force from the linear-comb-drive actuator displaces not only the comb drive and the folded spring that are connected to it but also the rotary hinges in the parallelogram four-bar linkage. In order to correctly predict and design the displacement in the XY plane, the relationship between the end-effector displacement and the angular displacement of the hinges in the four-bar mechanism has to be derived. Equation (3) gives such a relationship between the displacement of the end effector and the rotation angle of four-bar linkages. Since the angular deflection of the hinges and the displacement of the table are relatively small when compared to the overall dimension of the mechanism, the relationship between the angular deflection of the hinges in four-bar linkages and the displacement of the table can be linearized with Jacobian matrix $J2$ and its inverse $J2_{\mathrm{inv}}$

$$\begin{bmatrix}\Delta\theta_x\\ \Delta\theta_y\end{bmatrix} = J2_{\mathrm{inv}}\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix}, \qquad \text{where}$$

$$J2_{\mathrm{inv}} = \begin{bmatrix} \frac{\delta\theta_x}{\delta x}\Big|_{(-d_0-L,d_0+L)} & \frac{\delta\theta_x}{\delta y}\Big|_{(-d_0-L,d_0+L)} \\ \frac{\delta\theta_y}{\delta x}\Big|_{(-d_0-L,d_0+L)} & \frac{\delta\theta_y}{\delta y}\Big|_{(-d_0-L,d_0+L)} \end{bmatrix}. \tag{12}$$

Partially differentiate (3) with respect to x, y

$$\frac{\delta\theta_x}{\delta x} = 0 \quad \frac{\delta\theta_x}{\delta y} = -\frac{1}{L\sqrt{1-\left(\frac{d_0+L-y}{L}\right)^2}}$$

$$\frac{\delta\theta_y}{\delta x} = \frac{1}{L\sqrt{1-\left(\frac{x+d_0+L}{L}\right)^2}} \quad \frac{\delta\theta_y}{\delta y} = 0. \tag{13}$$

Evaluating (12) at $T_0(-d_0 - L, d_0 + L)$, we obtain $\delta\theta_x/\delta x|_{T_0} = 0$, $\delta\theta_x/\delta y|_{T_0} = -(1/L)$, and $\delta\theta_y/\delta x|_{T_0} = (1/L)$, $\delta\theta_y/\delta y|_{T_0} = 0$. As a result, inverse Jacobian matrix $J2_{\mathrm{inv}}$ is given by

$$J2_{\mathrm{inv}} = \begin{bmatrix} 0 & -\frac{1}{L} \\ \frac{1}{L} & 0 \end{bmatrix}. \tag{14}$$

In our stage design, the value of parameter L is 1 mm, so mapping of the stage displacement to the hinge angular displacement is given by $J2_{\mathrm{inv}} = \begin{bmatrix} 0 & -1000 \\ 1000 & 0 \end{bmatrix}$.

The relation between the rotary deflections of the hinges in the four-bar mechanism and the linear displacements from prismatic joints (linear comb drives) can be derived by combining (4) and (12)

$$\begin{aligned}\begin{bmatrix}\Delta\theta_x\\ \Delta\theta_y\end{bmatrix} &= J2_{\mathrm{inv}}\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix} = J2_{\mathrm{inv}}J\begin{bmatrix}\Delta d_x\\ \Delta d_y\end{bmatrix} \\ &= \begin{bmatrix} 0 & -\frac{1}{L} \\ -\frac{1}{L} & 0 \end{bmatrix}\begin{bmatrix}\Delta d_x\\ \Delta d_y\end{bmatrix} \\ &= \begin{bmatrix} 0 & -1000 \\ -1000 & 0 \end{bmatrix}\begin{bmatrix}\Delta d_x\\ \Delta d_y\end{bmatrix}. \end{aligned} \tag{15}$$

When comb drives are actuated, the work done by the comb-drive actuator is balanced by the energy stored in the hinges and the leaf springs. Using the principle of virtual work

$$\begin{aligned}&\frac{1}{2}F_{\mathrm{comb}-x}\Delta d_x + \frac{1}{2}F_{\mathrm{comb}-y}\Delta d_y \\ &\quad = \frac{1}{2}K_{\mathrm{spring}}(\Delta d_x)^2 + \frac{1}{2}K_{\mathrm{spring}}(\Delta d_y)^2 + \frac{1}{2}\times 4 \\ &\qquad \times K_{\mathrm{hinge}}(\Delta\theta_x)^2 + \frac{1}{2}\times 4 \times K_{\mathrm{hinge}}(\Delta\theta_y)^2. \end{aligned} \tag{16}$$

The equation can be expressed in matrix form

$$\begin{aligned}&\begin{bmatrix}F_{\mathrm{comb}-x}\Delta d_x\\ F_{\mathrm{comb}-y}\Delta d_y\end{bmatrix} \\ &\quad = \left\{\begin{bmatrix} K_{\mathrm{spring}} & 0 \\ 0 & K_{\mathrm{spring}} \end{bmatrix} \right. \\ &\qquad \left. + 4\begin{bmatrix} K_{\mathrm{hinge}} & 0 \\ 0 & K_{\mathrm{hinge}} \end{bmatrix}(J2_{\mathrm{inv}}J)^2\right\}\begin{bmatrix}\Delta d_x\\ \Delta d_y\end{bmatrix}^2 \end{aligned} \tag{16.1}$$

where $F_{\mathrm{comb}-x}$ and $F_{\mathrm{comb}-y}$ are the actuation forces for the X- and Y-axes, respectively; K_{spring} is the stiffness of the folded spring in the displacement direction; and K_{hinge} is the rotary stiffness of the flexure hinge. The coefficient 4 before K_{hinge} comes from four identical hinges inside a four-bar mechanism, where the four hinges are deflected equally when actuated.

Assuming that only one axis (X) is actuated, we have $F_{\mathrm{comb}-x}\Delta d_x = K_{\mathrm{spring}}\Delta d_x^2 + 4K_{\mathrm{hinge}}(1/L)^2\Delta d_x^2$, so

$$\Delta d_x = \frac{F_{\mathrm{comb}-x}}{K_{\mathrm{spring}} + 4K_{\mathrm{hinge}}/L^2}. \tag{17}$$

From (17), using the designed parameters, Δd_x is calculated as 14.5 μm at an actuation voltage of 150 V. Due to symmetry in the design, the same displacement can be achieved for the Y-axis.

IV. Dynamic Analysis and FEA Results

The natural frequency and modal shapes of the designed device are important design parameters for its successful

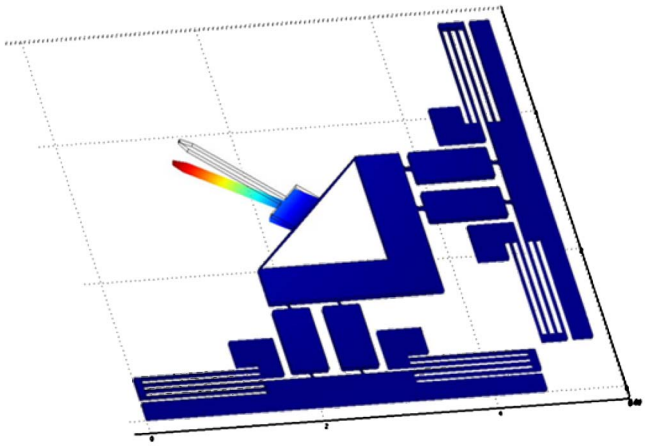


Fig. 4. Device deflection under self-weight and a load on the cantilever plate.

application to the high-bandwidth motion required in high-throughput nanomanipulation and nanomanufacturing. A high designed natural frequency enables the device to response quickly and accurately to the rapid changes in the commanded signals. In order to verify our device design, a finite-element analysis (FEA) (using COMSOL) is performed to study the dynamic behavior of the structure and to estimate the resonant frequencies and the associated mode shapes. A 3-D model is generated by AutoCAD and processed with COMSOL. Due to node and memory limitations of the available FEA simulation software, the fine triangular configuration of the moving parts of the stage (which are used to reduce mass) is replaced by a solid geometry. To compensate for the extra mass introduced by this approximation, the material density of the moving parts is scaled down appropriately.

Out-of-plane sagging of the stage and the cantilever due to its own weight can cause problems in our design by reducing the gap of the parallel-plate actuator and, consequently, the out-of-plane motion. Furthermore, it can lead to the twisting of the leaf springs and misalignment between the comb fingers, affect the orientation of the stage, and introduce additional stresses at the hinges. Thus, it is necessary to check the self-sagging of the stage and the effect of tilt-plate actuation and torsion bars on the stage through an FEA model. Two types of load are given for FEA. One is the surface load corresponding to the weight of the structure that is applied to the top surface of the device, and the other is a point load along the Z-direction acting at the center of the tilt-plate actuator to simulate the actuation force for the cantilever. The result of FEA is shown in Fig. 4. The sagging effect of the stage itself under the self-weight and actuation force of the cantilever is negligible (less than 5 nm) compared with the overall dimension of the end effector and the gap of the tilt-plate actuator. The out-of-plane cantilever deflection is mainly caused by deformation of the torsion bar. The interference between the vertical actuation structure and the lateral displacement structure is minimized and decoupled.

Since the structural responses in the lateral and vertical directions are decoupled, we decompose the eigen-frequency analysis into two subproblems, namely, lateral modes of the stage system and out-of plane mode of the torsion bar and the cantilever, to simplify the FEA problem and satisfy the node and memory limitations of the available FEA simulation software. Furthermore, while one might expect interaction between the lateral and vertical motions because of the torsion spring, as evident in subsequent discussions, this interaction takes place at a frequency that is much higher than the frequencies associated with the actuation modes in the lateral direction.

The natural frequency and mode shapes of the stage in the lateral direction are analyzed by FEA, and Fig. 5 shows the first six most dominant mode shapes for the system. The different color indicates the different displacement in the mode vibration. Among these modes, the first three modes are related to the displacement of the end effector (through the compliance of the folded springs and the parallelogram four-bar linkages), while the last three ones are related to the lateral dynamics of the torsion bars and the folded springs. Mode 1 has the leaf-spring deformations in antiphase, while mode 2 has them in phase. It can be seen that mode 3 is a rotational mode, while the first two modes are the translational modes. The two translational modes correspond to the in-phase and out-of-phase displacements of the two comb drives (and related flexure springs) and result in a displacement of the end effector in two perpendicular directions. The modal frequencies are slightly different because of the small differences in the end-effector displacements that occur when the four-bar systems rotate in the same directions or in different directions. Mode 3, in which the platform undergoes a rotation, which is an undesirable parasitic motion for this flexure mechanism, is roughly 16 times stiffer than the modes associated with the desired XY motions. This is attributed to the parallel kinematic XY stage design which, besides producing a relatively high natural frequency associated with the desired modal directions (the desired translational degree of freedom in the XY plane), also provides for good separation between the modes associated with the desired motion and those associated with the parasitic motion. The frequency separation can be even upgraded through the improved design of the folded leaf spring.

The modal frequency and corresponding mode shapes of the cantilever and torsion bars in the vertical directions are analyzed by only modeling the end effector of the stage and the cantilever structure. The end effector of the stage is assumed to be stationary. Fig. 6 shows the first four dominant mode shapes for the cantilever system. The color indicates the displacement in the mode vibration. Among these modes, the first mode is related to the out-of-plane rotation of the cantilever, while the second to fourth modes are related to the combined dynamics in the lateral direction and the out-of-plane direction of the torsion bars. The modes 2 and 3 in Fig. 6 are related to the modes 4 and 5 in Fig. 5. Due to the extra vibration component in the Z-direction, the frequencies of modes 2 and 3 in Fig. 6 are larger than that in Fig. 5 (in-plane modes). Again, the parasitic modes (modes 2–4) have much larger resonant frequencies and are located far from the first dominant modes (for both the lateral and out-of-plane motions), indicating a much higher stiffness to excite these parasitic motions. Our designed response for the cantilever structure is out-of-plane rotation associated with the dominant mode.

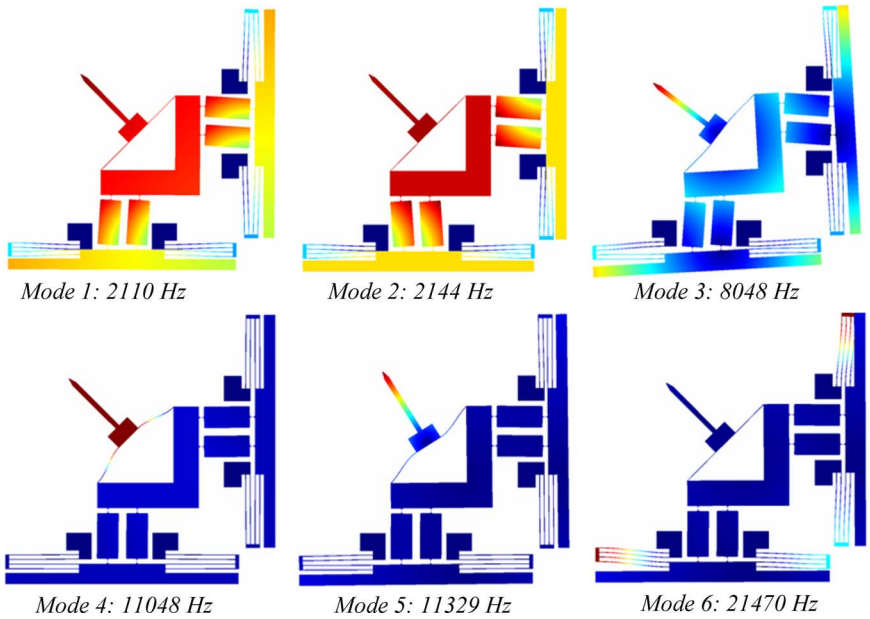


Fig. 5. Lateral modal shapes and their corresponding natural frequencies.

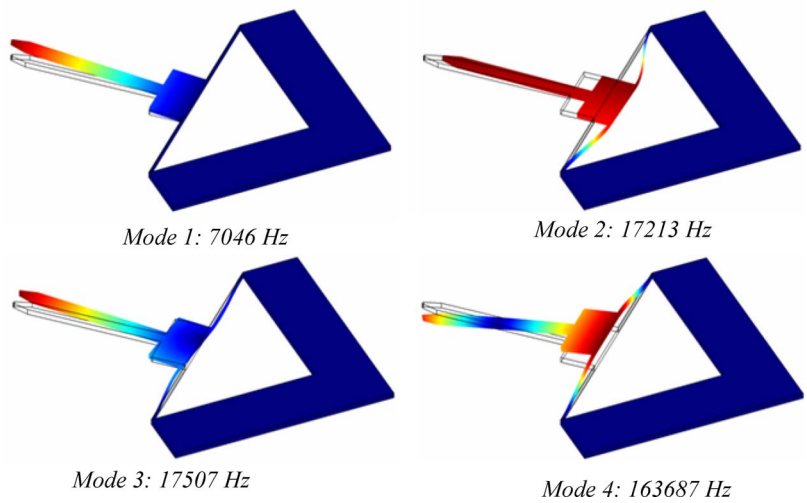


Fig. 6. Modal shapes and their corresponding natural frequencies of the cantilever.

V. Fabrication

The microfabrication processes that are utilized to fabricate the active cantilever device with micropositioning XY stage are shown in Fig. 7. The processes use four photolithography masks for transferring patterns of the device, i.e., three for device-side patterning and one for handle-layer patterning. Three device-side patterns consist of a silicon nitride pattern used to prestress the cantilever, a conducting Au/Cr pad pattern for future wire bonding, and a device pattern for defining structural components, such as the comb-drive actuators, the cantilever, the torsion bars, etc. The actuators and mechanism are designed to permit a monolithic fabrication of the structure. The mask for the handle layer defines one electrode for the tilt-plate actuator for the cantilever. The device is fabricated on a silicon-on-insulator (SOI) wafer with a device layer thickness of 50 μm and a buried oxide (BOX) layer thickness of 3 μm, which is supported on a 600-μm-thick handle layer. The device with an overall size of 4 mm $\times$ 4 mm of the bounding box without the contact pads is fabricated on a die with a 15-mm $\times$ 15-mm entire size. The device that is fabricated on an SOI substrate allows for the two parts (stators) of the electrostatic linear comb drives (fabricated on the device layer), and one electrode of the tilt-plate actuator (fabricated on the handle layer), to be electrically isolated from each other (by the BOX layer), while the device layer components are structurally supported by the handle layer.

The microfabrication process begins with the SC-1-cleaned 15-mm-square pieces of the diced wafer. The method of standard cleaning 1 (SC-1, $100:10:1$ of $H_2O:H_2O_2:NH_4OH$) is performed to remove the debris and clean the surface of the die after dicing [Fig. 7(a)]. A silicon nitride layer that is used to prestress the cantilever for initial out-of-plane bending is deposited on the device side using plasma-enhanced chemical vapor deposition for creating a high-stress film. This is followed

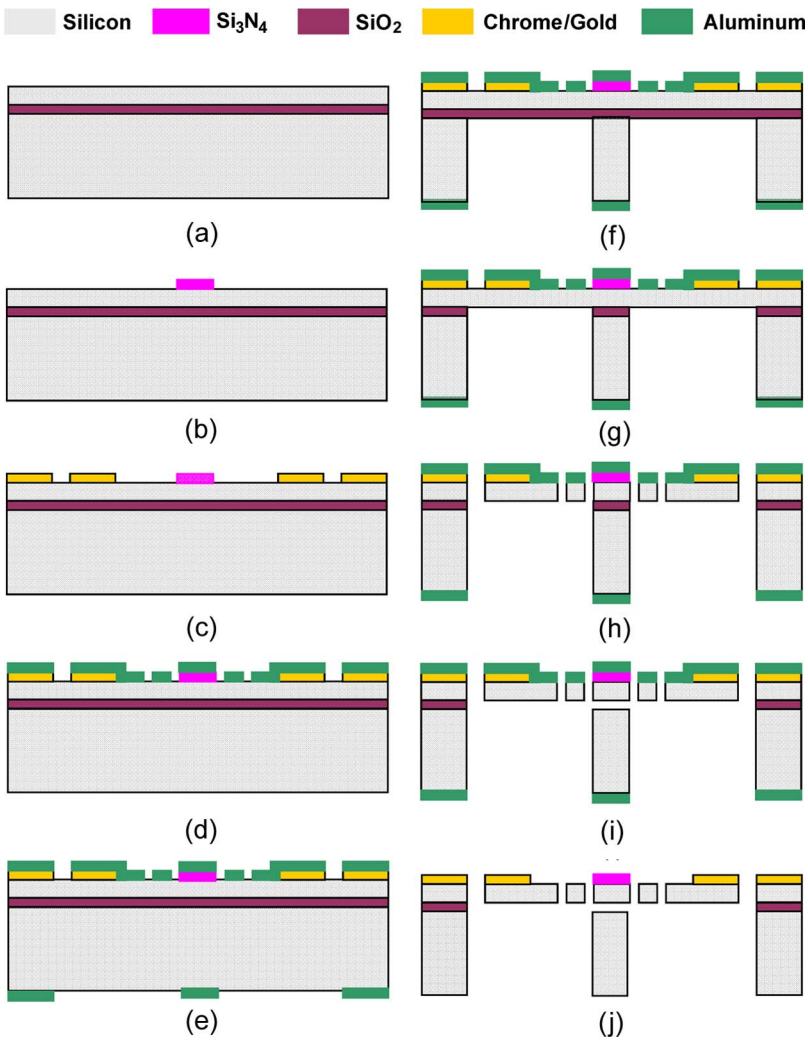


Fig. 7. Process flow for fabricating the active cantilever device with XY stage. (a) Initial SOI substrate. (b) Silicon nitride deposition and patterning to form a prestress layer for the cantilever. (c) Chrome/Gold deposition to form the pad layers. (d) Front-side aluminum mask patterning. (e) Back-side aluminum mask patterning. (f) DRIE of the back side. (g) Removal of the BOX layer by concentrated HF. (h) DRIE etching to release the device. (i) Releasing of parallel plate for the cantilever by HF etching. (j) Removal of aluminum mask by aluminum etchant.

by photolithographic patterning with photoresist AZ 4620 and by reactive ion etching of the exposed silicon nitride using CF_4 plasma [Fig. 7(b)]. The electrical contact pads that are used for wire bonding are composed of a chrome (17-nm) and gold (392-nm) stack. These contact pads are aligned with silicon nitride pattern and fabricated on the device layer by a metal sputtering step, followed by patterning of the pad layer with photoresist AZ 1518 manufactured by AZ Electronic Materials, and wet chemical etching of the exposed metals by gold etchants (3 min) and chrome etchants (1 min), respectively [Fig. 7(c)]. After that, the device pattern, including all the mechanisms and actuators, is aligned with the contact pads and patterned on the device layer of the SOI die by photolithography using photoresist AZ 1518. The final pattern is achieved by sputtering of a 60-nm-thick aluminum layer and lifting off of aluminum in an acetone bath through ultrasonic cleaning [Fig. 7(d)]. The device layer is then protected by spin coating and hard baking a thin layer of photoresist (5-μm-thick AZ1518) to protect the device-side pattern for the following fabrication steps. The die is flipped over, and the handle-layer pattern that defines the electrode for actuating the cantilever is aligned with the device-layer pattern and fabricated by the same process used for patterning the device layer [Fig. 7(e)]. Next, the deep reactive ionic etching (DRIE) with Bosch process using the STS Multiplex Advanced Silicon Etcher equipment is used to etch the handle layer from the back of the device so as to make the electrode of the tilt-plate actuator. The etching cycle and passivation cycle time of the BOSCH processes are optimized to yield a smooth sidewall profile with high aspect ratio [Fig. 7(f)]. The exposed BOX layer was etched by using concentrated HF (49%) acid [Fig. 7(g)]. The location of this step in the fabrication step sequence is crucial; otherwise, the residual stresses from the silicon dioxide film can destroy the device during the subsequent DRIE step. Following the box layer removal step, the device side of the die is subjected to the DRIE Bosch process to etch the device pattern through the device layer, leaving the different parts of the electrostatic drives physically isolated from each other [Fig. 7(h)]. At the end of this step, the tilt plates of the capacitor for actuating the cantilever are still connected by the BOX layer.

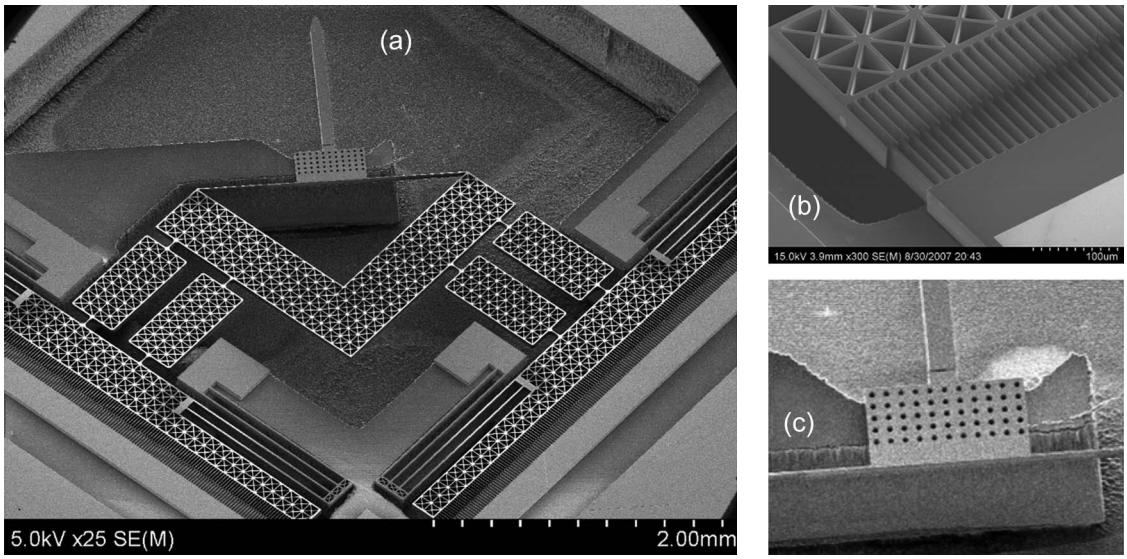


Fig. 8. SEM images of the fabricated device. (a) Overall structure. (b) Comb actuator for the stage. (c) Parallel-plate actuator for the cantilever.

They are released with the assistance of the release holes etched into the plate on the device layer plate by a vapor-phase HF release step that lasts for about 2.5 h [Fig. 7(i)]. In the last step, the Al films that served as masks for the Bosch process are removed by using aluminum etchant. The fabricated device is shown in Fig. 8.

VI. Characterization and Experimental Results

The fabricated device is characterized by supplying different driving voltages to the actuators (tilt-plate actuator for the cantilever and comb-drive actuators for the XY stage) and measuring the resulted displacement of the moving part of the actuators and the table. The modal frequency corresponding to the first mode of the XY stage is also measured and compared with the theoretically predicted value. The Q-factor associated with the stage is measured and reported. For static displacement characterization, a voltage supply (Keithley Model 237) is used to actuate the tilt-plate actuator and the comb drives of the stage, which are mounted and connected on a probe station.

For active cantilever characterization, a Veeco NT1000 noncontact-interferometry-based optical profiler is used to measure the static out-of-plane displacement of the cantilever. The measuring region is located at the center of the plate of the cantilever. The fixed plate that is located on the handle layer is used as the reference plane for detecting the vertical motion of the cantilever. We choose to measure at the plate of the cantilever instead of the tip of the cantilever because of the limited measuring area available at the tip and the difficulty of finding a reference plane near the tip of the cantilever. Fig. 9 shows the static displacement of the cantilever at the center of the tilt-plate actuator. The pull-in voltage is about 4.6 V with a maximum stable displacement of about 0.85 μm, which is very close to the analytical prediction. The vertical displacement of the stage that holds the torsion bar and the cantilever is measured to be negligible within the measuring resolution. The lever-arm effect should result in about 7-μm out-of-plane displacement at the tip of the cantilever.

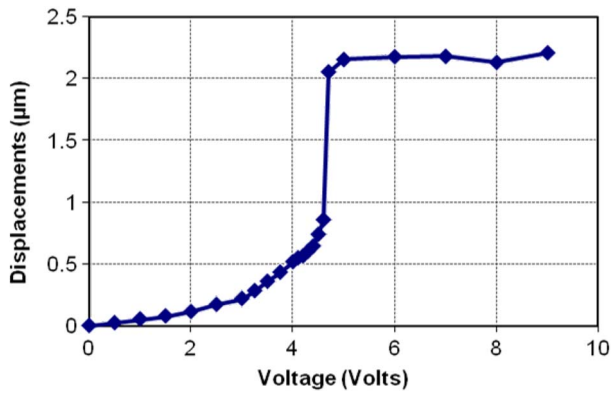


Fig. 9. Static displacement characterization of the active cantilever.

For the static displacement characterization of the micropositioning XY stage, a voltage is applied to the linear comb drive in increments of 10 V, and the resultant displacement is measured using a high-resolution microscope with a resolution of 1 μm attached to the probe station. Fig. 10(a) shows the displacement of the comb drive that is aligned with the X-axis of the stage as a function of actuation voltage. The comb drive moves by 24 μm under an actuation voltage of 180 V. Additionally, Fig. 10(b) shows that the displacement of the stage is linearly related to the square of the actuation voltage, as predicted by the electrostatic actuation theory. The other comb has almost identical static displacement curves. The displacements of the stage at the different voltages were found to be repeatable within the resolution of the optical microscope. Additionally, the maximum displacement of the stage is much less than the elastic working range of all the deforming elements versus the folded springs and flexure hinges. Hence, there is negligible material hysteresis of the complaint components of the stage mechanism. The displacement of the end effector can be reproduced accurately and repeatedly.

The dynamic behavior (natural frequency) of the stage is tested experimentally. A signal generator (HP/Agilent 33220A)

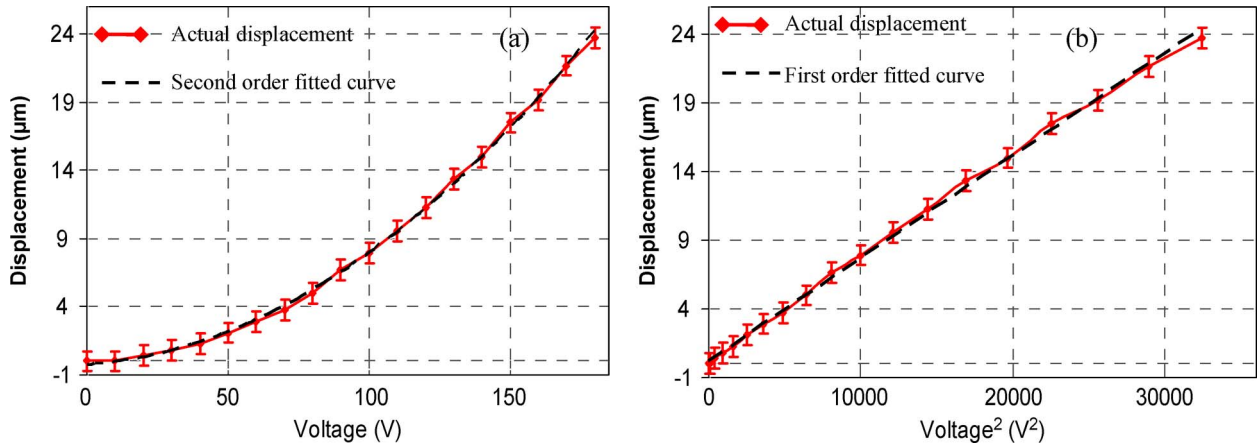


Fig. 10. Static displacement characterization of the XY stage. (a) Axis displacement as a function of actuation voltage. (b) Axis displacement varying linearly with the square of the actuation voltage.

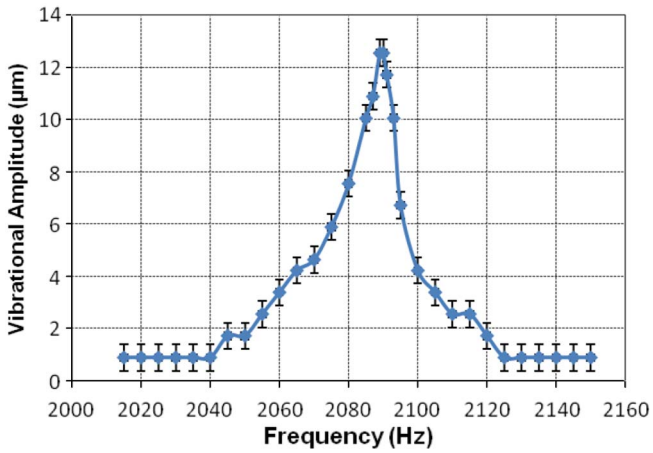


Fig. 11. Magnitude of vibration around the resonant frequency in air.

is used to generate sinusoid voltages with the required frequencies that are amplified by a voltage amplifier (Trek Model 623B) to 0–30 V peak-to-peak ($V = 15 + 15 \sin \omega t$). This output, with varying frequencies, is used to actuate the comb-drive actuators. The resulting vibrational amplitudes are optically recorded through the microscope. For each sampled frequency, the steady-state response that is the vibration amplitude is measured optically with a resolution of 1 μm. After an initial sweep through the frequencies, the device's resonant frequency is located. Careful measurements are then made in a smaller frequency window around the device's resonant frequency to obtain the frequency response shown in Fig. 11. Clearly, the resonant frequency of the device under test is about 2090 Hz, while the theoretical prediction from FEA is about 2113 Hz. The discrepancy comes from the dimensional variation from the fabrication process. Small changes in the dimensions between the design and the actual device affect the stiffness of the fabricated structures and, thus, the displacement and natural frequency. The frequencies with amplitudes equal to $1/\sqrt{2}$ of the maximum peak are around 2084 and 2094 Hz. Thus, the Q-factor is approximately equal to $Q = (f_0/\Delta f) = 2090/10 \approx 210$ in air. This relatively high Q-factor is attributed to the high stiffness and good modal separation that result from the parallel kinematic stage design. The large modal separation avoids the superposition of the neighboring resonant peak. The superposed resonant peaks may enlarge the effective frequency band of the resonant peak Δf and decrease the Q-factor. The other reason for the high quality factor is the etching away of the handle layer, which decreases the film damping between the stage and the handle wafer.

VII. Conclusion

In this paper, an active cantilever device integrated with a high-bandwidth nanopositioning XY stage that was based on a parallel kinematic mechanism was designed, analyzed, fabricated, and characterized. The cantilever was connected to the end effector of the XY stage through two torsion bars that provided the rotary compliance of the cantilever. The cantilever was actuated electrostatically through a separate electrode that was fabricated beneath the cantilever. The active cantilever as a functional manipulator was carried by a micropositioning stage, which enabled high-bandwidth scanning and manipulation in three dimensions. The parallel-kinematic-based micropositioning XY stage design produced linear kinematics in the operating motion range of the stage and generated well-dispersed modal frequencies, with the dominant modes being the desired translations. The relatively simple kinematics and dynamics facilitated future control design for a closed-loop positioning system. FEA simulations verified the design objectives. The integrated cantilever device was fabricated on an SOI die, and high-aspect-ratio features were fabricated by using deep reactive ion etching (DRIE) processes. The actuation electrode of the cantilever was fabricated on the handle layer, while the cantilever and XY stage were at the device layer of the SOI wafer. Experimental testing suggested that an estimated 7-μm out-of-plane motion of the cantilever tip was obtained at 4.5 V, and an observed 24 μm of lateral stage motion was achieved at 180 V. The dominant natural frequency of the stage system was measured to be about 2090 Hz. A high Q-factor ($\sim$200) was achieved due to the high-stiffness parallel kinematic design. The fabricated stages were adapted as chip-scale manufacturing and metrology tools for nanomanufacturing and nanometrology applications.

REFERENCES

[1] G. Binnig, H. Rohrer, C. Gerber, and E. Weibel, "Tunneling through a controllable vacuum gap," *Appl. Phys. Lett.*, vol. 40, no. 2, pp. 178–180, Jan. 1982.

[2] G. Binnig, H. Rohrer, C. Gerber, and E. Weibel, "Surface studies by scanning tunneling microscopy," *Phys. Rev. Lett.*, vol. 49, no. 1, pp. 57–61, Jul. 1982.

[3] G. Binnig, C. F. Quate, and C. Gerber, "Atomic force microscope," *Phys. Rev. Lett.*, vol. 56, no. 9, pp. 930–933, Mar. 1986.

[4] B. Bhushan, *Handbook of Micro/Nano Tribology*, 2nd ed. Boca Raton, FL: CRC Press, 1999.

[5] E. Meyer, H. J. Hug, and R. Bennewitz, *Scanning Probe Microscopy: The Lab on a Tip*, 1st ed. Berlin, Germany: Springer-Verlag, 2003.

[6] Y. Lee, G. Lim, and W. Moon, "A piezoelectric micro-cantilever biosensor using the mass-micro-balancing technique with self-excitation," *Microsyst. Technol.*, vol. 13, no. 5/6, pp. 563–567, Mar. 2007.

[7] F. M. Battiston, J. P. Ramseyer, H. P. Lang, M. K. Baller, C. Gerbe, J. K. Gimzewski, E. Meyer, and H. J. Güntherodt, "A chemical sensor based on a microfabricated cantilever array with simultaneous resonance-frequency and bending readout," *Sens. Actuators B, Chem.*, vol. 77, no. 1/2, pp. 122–131, Jun. 2001.

[8] R. D. Piner, J. Zhu, F. Xu, S. Hong, and C. A. Mirkin, "Dip pen nanolithography," *Science*, vol. 283, no. 5402, pp. 661–663, 1999.

[9] P. Vettiger, M. Despont, U. Drechsler, U. Durig, W. Haberle, M. I. Lutwyche, H. E. Rothuizen, R. Stutz, R. Widmer, and G. K. Binnig, "The 'Millipede'—More than one thousand tips for future AFM data storage," *IBM J. Res. Develop.*, vol. 44, no. 3, pp. 323–340, May 2000.

[10] W. P. King, T. W. Kenny, K. E. Goodson, G. L. W. Cross, M. Despont, U. Dürig, H. Rothuizen, G. Binnig, and P. Vettiger, "Atomic force microscope cantilevers for combined thermomechanical data writing and reading," *Appl. Phys. Lett.*, vol. 78, no. 9, pp. 1300–1302, Feb. 2001.

[11] C. F. Quate, "Scanning probes as a lithography tool for nanostructures," *Surf. Sci.*, vol. 386, no. 1–3, pp. 259–264, Oct. 1997.

[12] X. Wang, D. A. Bullen, J. Zou, C. Liu, and C. A. Mirkin, "Thermally actuated probe array for parallel dip-pen nanolithography," *J. Vac. Sci. Technol. B, Microelectron. Nanometer Struct.*, vol. 22, no. 6, pp. 2563–2567, 2004.

[13] D. Lee, T. Ono, and M. Esashi, "High-speed imaging by electromagnetic alloy actuated probe with dual spring," *J. Microelectromech. Syst.*, vol. 9, no. 4, pp. 419–4124, Dec. 2000.

[14] J. H. Park, T. Y. Kwon, D. S. Yoon, H. Kim, and T. S. Ki, "Fabrication of microcantilever sensors actuated by piezoelectric $Pb(Zr_{0.52}Ti_{0.48})O_3$ thick films and determination of their electromechanical characteristics," *Adv. Funct. Mater.*, vol. 15, no. 12, pp. 2021–2028, 2005.

[15] S. S. Lee and R. M. White, "Self-excited piezoelectric cantilever oscillators," *Sens. Actuators A, Phys.*, vol. 52, no. 1–3, pp. 41–45, Mar./Apr. 1996.

[16] W. M. Huang, Q. Y. Liu, L. M. He, and J. H. Yeo, "Micro NiTi–Si cantilever with three stable positions," *Sens. Actuators A, Phys.*, vol. 114, no. 1, pp. 118–122, Aug. 2004.

[17] D. A. Bullen and C. Liu, "Electrostatically actuated dip pen nanolithography probe arrays," *Sens. Actuators A, Phys.*, vol. 125, no. 2, pp. 504–511, Jan. 2006.

[18] V. P. Jaecklin, C. Linder, N. F. de Rooij, J. M. Moret, R. Bischof, and F. Rudolf, "Novel polysilicon comb actuators for XY-stages," in *Proc. IEEE Micro Electro Mech. Syst. Workshop*, 1992, pp. 147–149.

[19] C. S. B. Lee, S. Han, and N. C. MacDonald, "Single crystal silicon (SCS) XY-stage fabricated by DRIE and IR alignment," in *Proc. IEEE MEMS*, 2000, pp. 28–33.

[20] C. H. Kim and Y. K. Kim, "Micro XY-stage using silicon on a glass substrate," *J. Micromech. Microeng.*, vol. 12, no. 2, pp. 103–107, Feb. 2002.

[21] P. F. Indermuhle, V. P. Jaecklin, J. Brugger, C. Linder, N. F. de Rooij, and M. Binggeli, "AFM imaging with an xy-micropositioner with integrated tip," *Sens. Actuators A, Phys.*, vol. 47, no. 1–3, pt. 4, pp. 562–565, Mar./Apr. 1995.

[22] C. H. Kim, H. M. Jeong, J. U. Jeon, and Y. K. Kim, "Silicon micro XY-stage with a large area shuttle and no-etching holes for SPM-based data storage," *J. Microelectromech. Syst.*, vol. 12, no. 4, pp. 470–478, Aug. 2003.

[23] T. Harness and R. A. Syms, "Characteristic modes of electrostatic comb-drive $X-Y$ microactuators," *J. Micromech. Microeng.*, vol. 10, no. 1, pp. 7–14, Mar. 2000.

[24] K. Takahashi, H. N. Kwon, K. Saruta, M. Mita, H. Fujita, and H. Toshiyoshi, "A two-dimensional $f-\theta$ micro optical lens scanner with electrostatic comb-drive XY-stage," *IEICE Electron. Exp.*, vol. 2, no. 21, pp. 542–547, 2005.

[25] K. Takahashi, M. Mita, H. Fujita, and H. Toshiyoshi, "A high fill-factor comb-driven XY-stage with topological layer switch architecture," *IEICE Electron. Exp.*, vol. 3, no. 9, pp. 197–202, 2006.

[26] J. Dong, D. Mukhopadhyay, and P. M. Ferreira, "Design, fabrication and testing of silicon-on-insulator (SOI) MEMS parallel kinematics XY stage," *J. Micromech. Microeng.*, vol. 17, no. 6, pp. 1154–1161, May 2007.

[27] D. Mukhopadhyay, J. Dong, E. Pengwang, and P. M. Ferreira, "A SOI-MEMS-based 3-DOF planar parallel-kinematics nanopositioning stage," *Sens. Actuators A, Phys.*, vol. 147, no. 1, pp. 340–351, Sep. 2008.

[28] D. Mukhopadhyay, J. Dong, and P. M. Ferreira, "Parallel-kinematic-mechanism-based monolithic XY micropositioning stage," in *Proc. SPIE Photonics West, MOEMS-MEMS Conf.*, San Jose, CA. Jan. 2008, pp. 688 209-1–688 209-7.

[29] Q. Yao, J. Dong, and P. M. Ferreira, "Design, analysis, fabrication and testing of a parallel-kinematic micropositioning XY stage," *Int. J. Mach. Tools Manuf.*, vol. 47, no. 6, pp. 946–961, May 2007.

[30] Q. Yao, J. Dong, and P. M. Ferreira, "A novel parallel-kinematics mechanisms for integrated, multi-axis nanopositioning: Part 1. Kinematics and design for fabrications," *Precis. Eng.*, vol. 32, no. 1, pp. 7–19, Jan. 2008.

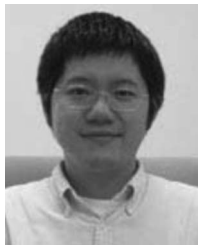
[31] J. Dong, Q. Yao, and P. M. Ferreira, "A novel parallel-kinematics mechanism for integrated, multi-axis nanopositioning: Part 2: Dynamics, control and performance analysis," *Precis. Eng.*, vol. 32, no. 1, pp. 20–33, Jan. 2008.

[32] J. Dong and P. M. Ferreira, "Simultaneous actuation and displacement sensing for electrostatic drives," *J. Micromech. Microeng.*, vol. 18, no. 3, p. 035 011, Jan. 2008. (10 pp).

[33] M. Bao, *Analysis and Design Principles of MEMS Devices*. Amsterdam, The Netherlands: Elsevier, Jun. 10, 2005.

[34] J. M. Paros and L. Weisbord, "How to design flexure hinges," *Mach. Des.*, vol. 37, pp. 151–156, Nov. 1965.

[35] R. Legtenberg, A. W. Groeneveld, and M. Elwenspoek, "Comb-drive actuators for large displacements," *J. Micromech. Microeng.*, vol. 6, no. 3, pp. 320–329, Sep. 1996.



Jingyan Dong received the B.S. degree in automatic control from the University of Science and Technology of China, Hefei, China, in 1998, the M.S. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2001, and the Ph.D. degree in mechanical engineering from the University of Illinois at Urbana-Champaign in 2006.

From 2006 to 2008, he was a Postdoctoral Research Associate at the Center for Nanoscale Chemical-Electrical-Mechanical Manufacturing Systems (Nano-CEMMS) at the University of Illinois at Urbana-Champaign. Since 2008, he has been a faculty member in the Department of Industrial and Systems Engineering, North Carolina State University, Raleigh. His research interests include micro/nano manufacturing, MEMS, multidimensional-scale mechatronics and manufacturing systems, and their design, fabrication, sensing, control, and application to micro/nano/macro manipulation and manufacturing.

Dr. Dong is a member of the American Society of Mechanical Engineers (ASME).

Placid M. Ferreira received the Ph.D. degree in industrial engineering from Purdue University, West Lafayette, IN, in 1987.

He is currently the Grayce Wicall Gauthier Professor of Mechanical Science and Engineering at the University of Illinois at Urbana-Champaign, where he directs the NSF-sponsored Center for Nanoscale Chemical-Electrical-Mechanical Manufacturing Systems (Nano-CEMMS). His research interests include precision engineering, automation and control of manufacturing systems, MEMS, and nanoscale manufacturing.

Prof. Ferreira is a member of the American Society of Mechanical Engineers (ASME).

IOP PUBLISHING MEASUREMENT SCIENCE AND TECHNOLOGY

Meas. Sci. Technol. **19** (2008) 045203 (13pp) doi:10.1088/0957-0233/19/4/045203

Self-calibration of dual-actuated single-axis nanopositioners

Y H Jeong, J Dong and P M Ferreira

Department of Mechanical Science and Engineering and Center for Nanoscale Chemical-Electrical-Mechanical Manufacturing Systems, University of Illinois at Urbana-Champaign, 1206 W Green Street, Urbana, IL 61801, USA

E-mail: yhjeong@uiuc.edu, jdong@uiuc.edu, pferreir@uiuc.edu

Received 17 August 2007, in final form 21 December 2007
Published 13 February 2008
Online at stacks.iop.org/MST/19/045203

Abstract

The rapid growth in nanosciences and technology has increased the need for high-precision nanopositioning stage technology, an important aspect of which is calibration to a traceable length standard with nanometre resolution. Direct calibration of the entire displacement range to a traceable length standard is generally difficult and time consuming because of the dearth of suitable and stable references and the need to remove all environmental disturbances during the calibration procedure. This paper introduces an approach to implementing self-calibration on a single-axis, dual-actuated, nanopositioning stage. It demonstrates how dual actuation on such a system can be used to implement the transitivity and redundancy conditions required for dimensional self-calibration so that only a single scaling input is required. The approach is verified by a series of simulations and experiments to demonstrate repeatable self-calibration of the axis to within 1 nm over a displacement range of 30 μm.

Keywords: self-calibration, auto-calibration, nanopositioning, metrology, dual-stage actuation

1. Introduction

In a precision positioning system, the positioning accuracy and repeatability are possibly the two most important characteristics. Accuracy of a positioning system is related to the fidelity or conformity of the displacements it produces in some chosen reference system with an agreed-upon length standard, while the repeatability is the degree to which it repeats its response under similar displacement commands. Informally, if the stage's displacement response to repeated positioning commands is characterized by a distribution, the conformity of the mean of the distribution to the length standard [1] is a measure of its accuracy and the variance of the distribution is a measure of its repeatability. More detailed discussions may be found in [2, 3]. With the rapid growth of nanosciences and technology, the need for positioning stage technology with both accuracy and repeatability measured in the range of nanometres or fractions of a nanometre, while having displacement ranges several orders of magnitude larger, has increased considerably. For example, the image fidelity of a scanning probe and optical microscope (SPMs and NSOMs) depends, to a very large degree, on the accuracy of its scanning stage. Nanofabrication with electron and ion beams and many of the emerging (direct-write) processes also depend on the accuracy of the positioning stage for producing accurate structures. While traceability to a common length standard with nanoscale resolution is an immediate apparent requirement for metrology applications, the rather high size dependence of many phenomena that exploit nanoscale features and the emerging need for multiple processing steps on different tools make it an increasingly critical problem. Acknowledging the difficulty and challenges of obtaining and maintaining nanoscale repeatability, this paper addresses the problem of obtaining and maintaining nanoscale accuracy on a dual-actuated, single-axis nanopositioning stage.

The positioning accuracy of a stage deteriorates as a result of errors in assembly and inaccurate sensing systems. Slow changes in the errors occur as a result of 'quasistatic effects' such as flexural deflections, relaxation of internal stress in the members, slow changes in sensing gains/calibration, wear, flexural and thermal strains and other such effects [4, 5]. As a result, in spite of initial calibration, periodic calibration to known length standards is required to compensate for such systematic errors (or errors that persist and repeat as a function of position and other measurable quantities such as temperature, loading, etc) that enter a system [5–7]. Such

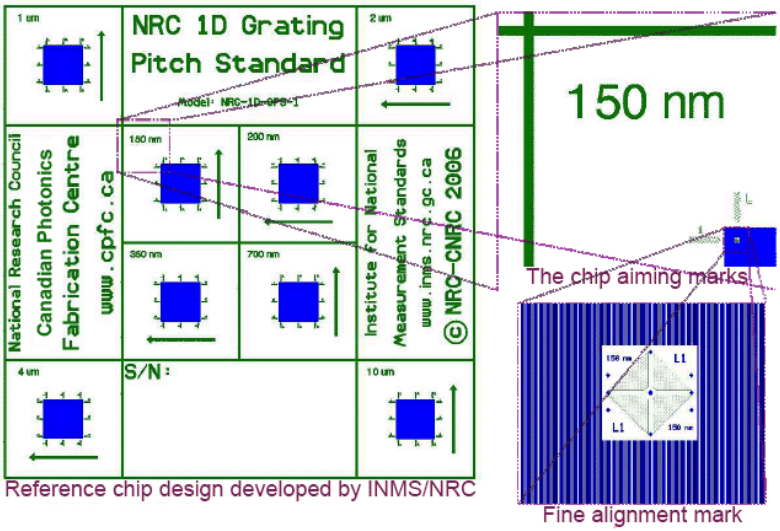


Figure 1. Different patterns for lateral calibration [14].
(This figure is in colour only in the electronic version)

calibrations are typically carried out by using a dimensional measurement system or artefact whose accuracy is at least an order of magnitude better than that expected from the calibrated system. In general, calibration procedures are difficult to perform and time consuming, requiring highly qualified personnel and specialized instrumentation. A number of calibration and error measurement procedures are standardized so as to formalize and make their results verifiable and transferable [8–10]. In the context of nanopositioning, the problem is further aggravated because in many cases the reference artefact or transducer has errors of roughly the same order of magnitude as the device to be calibrated, the system is highly susceptible to environmental disturbances and measurement access for external transducers without perturbing the system is difficult. While national standard facilities have stabilized laser interferometers capable of picometre-level accuracies and resolution [11], the transfer of length measures with nanometre-level accuracies has proved challenging. The current work in calibration of SPMs is instructive of the challenges encountered. For example, a common method for lateral calibrations is through the use of high-resolution grid patterns [12, 13]. Figure 1 taken from [14] depicts the different patterns for lateral calibration of such systems with different resolutions. Challenges in transferable length scales arise from the variations within the artefact and lack of feature definition (rounded corners, edges that are not sharp and slanted sidewalls) that give rise to uncertainty. These uncertainties are further confounded by variations in the tool itself (for example, geometry of the probe) that are difficult to deconvolute from the measurements.

An approach to mitigating the dependence on precalibrated artefacts is to use a self-calibration approach in which the errors of the tool and the artefact are simultaneously estimated. Evans *et al* [15] provide a very nice discussion on how reversal, redundancy (in measurements) and other similar ideas are used in estimating tool errors such as perpendicularity errors. Raugh [16, 17] has done extensive work on the theory and application of self-calibration in 2D (XY) systems, identifying and explaining the role of symmetry, transitivity and redundancy in self-calibration (will be discussed in more detail in the following section). In essence, self-calibration in its most basic form involves an artefact with measurement features spaced over a regular pattern (a grid, for example). When the artefact is placed on the stage and a measurement is made, the observed error is composed of both the artefact error at that point and the stage error at some specific point in its working range. The pattern remains invariant over some set of rigid body motions (e.g., a translation by the pitch of the grid or a rotation about the z-axis by a multiple of 90°); this provides a means of making measurements that permute the errors of the positioning systems at specific locations with those of the artefact at specific measurement features. The idea is to have a set of equations, such that all the errors (on the artefact and the desired points on the stage) have a transitive relationship, so that knowledge of one error value leads to solvability for all errors (unknowns in the system of equations). This knowledge is inserted into the system through a single traceable standard scale value. Thus, self-calibration reduces the amount of traceable standard information required for calibration. Of course, the results of the calibration depend on the robustness of the procedure to measure errors and it is here that redundancy and multiple measurements are used. In other literature relevant to self-calibration in this context, Raugh has applied self-calibration to the calibration of a two-dimensional positioning system for electron beam lithography [18, 19]. Takac has discussed the congruency and transitivity in the one-dimensional calibration case to obtain reproducibility of calibration [20]. Lee and Ferreira have extended self-calibration to triangulation- and trilateration-based calibration in 2D and 3D spaces using a one-dimensional transducer [21], and then experimentally verified the methods using a coordinate measuring machine (CMM) [22].

In the context of SPM lateral calibration and nanopositioning stage technology, self-calibration alleviates

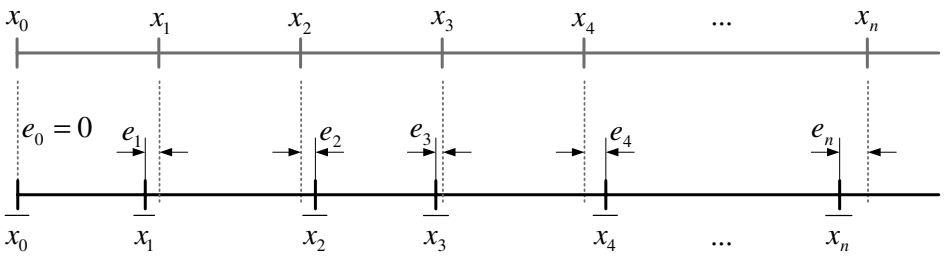


Figure 2. An ideal scale (grey) and a physical scale (black) with static error produce a correct calibration of the physical sensor by direct observation of differences.

the problem of inaccurate artefacts. However, it does not solve the problem of lack of feature definition. While the use of artefacts is convenient for devices that have measurement capabilities on them (metrology and imaging tools), it is quite difficult to measure artefacts with nanoscale resolution on stages being used for other purposes. Further, self-calibration processes that use such an external reference artefact are susceptible to errors in positioning and repositioning of the artefact (that are essential to produce the aforementioned equations that permute the stage and artefact errors).

In many nanotechnology applications, dual actuation is used to extend the range, resolution or bandwidth of the stage. For example, a piezo-actuated positioner may be mounted on a conventional leadscrew-driven stage to increase both the bandwidth and the resolution, or two piezo-actuated positioners may be stacked to increase the displacement range. The main contribution of this paper is to demonstrate how dual actuation of an axis can be exploited to make it self-calibrating. The procedure we develop does not require an external artefact. Each of the actuators plays the role of the artefact for calibrating the other. However, the procedure does require the introduction of standard scale information into the system (and this may be done by a simplified artifice with one traceably calibrated dimensional feature). Once introduced, this scale information can be passed along to all positions (scale graduation marks) on the axes through the transitive relationships developed by permuting displacements of the actuators to produce the same axis displacement. It turns out that the basic self-calibration procedure of Raugh [1] is fully applicable to this situation. Section 2 reviews the mathematical formulation of the self-calibration problem, while section 3 applies it to the dual-actuated, single-axis self-calibration problem. Section 4 gives simulation results for different error characteristics and measurement noise levels of each of the actuators/sensors of the two stages. It also discusses experiments involving a dual-actuated stage with nanometre resolution. Section 5 draws up conclusions and recommendations.

2. Self-calibration fundamentals

The displacement measured by a sensor can be divided into the following components: commanded or desired displacement, systematic or repeatable error and random, non-repeatable error or noise. The desired displacement or position is assumed to be known and measurable by a traceable length standard. Systematic error is a repeatable error that can be attributed to static and quasistatic effects such as assembly errors in the stage's structure, sensor inaccuracies, etc. The noise can be considered as random behaviour that accrues from external and internal disturbances such as electrical and environmental noise. Through proper control of the environmental conditions, electromagnetic shielding and other such arrangements, or through repetitions and subsequent averaging, these can be reduced to an acceptable level, but cannot be completely eliminated. Being random with no systematic information, such effects, grouped together, are assumed to be characterized by a normal distribution with a zero mean and a standard deviation reflecting the relative size of the uncertainty. The purpose of calibration (or self-calibration) is to assess the value of the systematic errors. As mentioned above, a displacement in a given reference frame and measured by the ith position of a scale can be expressed as follows:

$$M_i = T_i + Q_i + N(0, \sigma), \tag{1}$$

where M_i, T_i, Q_i and $N(0, \sigma)$ are the position measured by the scale, the desired or true position systematic error at the ith calibration position and the random error or measurement noise (normally distributed with a mean value of 0 and standard deviation of σ), respectively. In particular, since the formulation of the self-calibration problem only considers systematic errors (assuming that the effect of the noise term can be removed by repetitions and averaging), (1) can be simplified to

$$\overline{x_i} = x_i + e_i, \tag{2}$$

where $\overline{x_i}$, x_i and e_i are the position/displacement measured by scale, true position or displacement and the systematic error at the ith calibration position, respectively. Figure 2 schematically depicts the terms in this equation for one-dimensional or linear calibration. In figure 2, the grey grid or scale represents an ideal artefact or scale, while the black grid represents a physical scale with some error. In this case, the errors of the physical scale can be estimated by direct comparisons with the reference or ideal artefact. However, in general, an ideal reference artefact or scale is expensive to construct and difficult to maintain. A more common situation is to have the availability of two scales, both of which have errors. In such a situation, a measurement of error made at a position by comparing the two scales must consider the

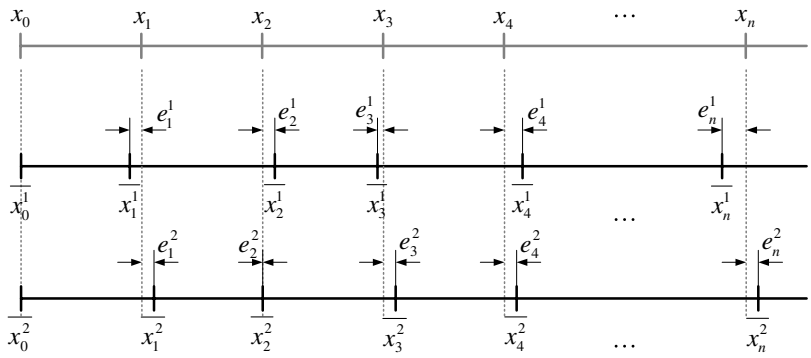


Figure 3. Two physical scales with their origins aligned [20].

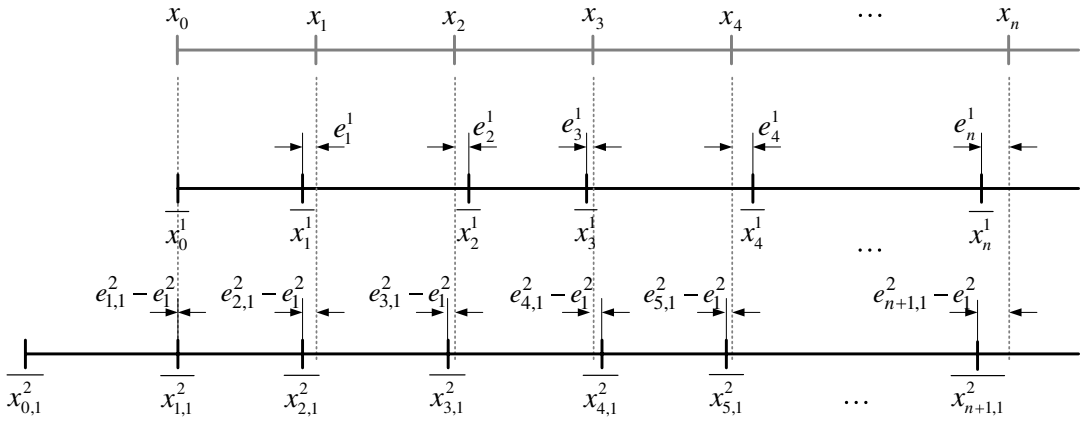


Figure 4. A comparison of two physical scales when the second scale is offset to align its first grid with the origin of the first scale [20].

errors inherent in both scales at that measurement position or displacement. The calibration problem then involves simultaneously assessing the errors of both scales and is called self-calibration. In the discussion that follows, the technique for resolving the errors of two inaccurate scales (one on the stage and the other being an external reference) is, in large part, adapted from [1, 15, 20].

Suppose we have two nominally identical measuring scales, each with n nominally equidistant graduation marks, but having different errors that we would like to estimate. Then, when the origins of the two scales are aligned, at each graduation or calibration point the difference in errors of the two scales can be written down as

$$\begin{aligned}\overline{x_i^1} - \overline{x_i^2} &= \left(x_i^1 + e_i^1\right) - \left(x_i^2 + e_i^2\right) \\ &= e_i^1 - e_i^2 = O_i, \qquad i = 1, 2, 3, \ldots, n, \end{aligned} \tag{3}$$

where $\overline{x_i^1}$, $\overline{x_i^2}$, x_i^1, x_i^2, e_i^1 and e_i^2 are the measured positions, desired or true positions and the systematic errors of the first and the second scales at the ith calibration point, respectively. Since the origins of the scales are aligned, the true positions for the two scales, x_i^1 and x_i^2, are the same and therefore cancel each other out in the equation. As shown in figure 3, which is a modified version of [20], by aligning the origins of two scales, the difference in error can be obtained from the difference in measurement at every point. However, the individual errors of the scales cannot be estimated from the difference because the number of errors to be estimated is twice the number of the measured differences. Moreover, the relationships do not satisfy transitivity requirements, making it infeasible for errors to be estimated, even with the input of a known value for one of the errors.

The problem of transitivity and the deficit in the number of equations can be solved by introducing an offset to the origin of the second scale. If the second scale is shifted by one graduation mark so that its first graduation mark is aligned with the origin of the first scale, then an observation of the differences in scale marks now involves the difference in errors at offset scale marks of the two scales. This provides the required transitivity to the set of equations, and it produces an additional n observations. Figure 4 shows the arrangement when such an offset is introduced between the scales. As shown in figure 4, the error pair composed in an observation is different. The second scale's error at the $(i+1)$th position is now involved with the ith position of the first scale in an observation. This new permutation of scale errors results in transitivity in the set of equations. Generalizing, if an offset involves aligning the jth graduation mark of the second scale with the origin of the first scale, the difference in measurement at the ith graduation point has the following relationship:

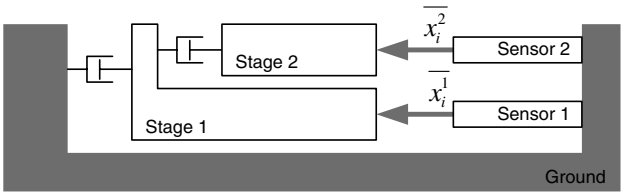


Figure 5. Configuration for the self-calibration of a dual-axis system.

$$\overline{x_i^1} - \left(\overline{x_{i,j}^2} - \overline{\text{Offset}_j}\right) = (x_i^1 + e_i^1) - \left\{\left(x_{i,j}^2 + e_{i,j}^2\right) - \left(x_j^2 + e_j^2\right)\right\}$$
$$= e_i^1 - e_{i+j}^2 + e_j^2 = O_i^j, \qquad j = 1, 2, 3, \ldots, m \tag{4}$$

where $\overline{x_i^1}$, x_i^1, e_i^1, $\overline{x_{i,j}^2}$, $x_{i,j}^2$, $e_{i,j}^2$, Offset_j and O_i^j are the measured, true positions and errors of the two scales when the offset Offset_j aligns the jth graduated position from the origin of the second scale with the origin of the first scale, and the measurement is being done at the ith graduated position of the first scale. Now consider the case when $j = 1$. The observation $O_i^1 = e_i^1 - e_{i+1}^2 + e_1^1$ when taken with the ith and (i+1)th equations of the system in (1), i.e. $O_i = e_i^1 - e_i^2$ and $O_{i+1} = e_{i+1}^1 - e_{i+1}^2$, produces the transitivity required between the error terms of the first scale, i.e. between e_i^1 and e_{i+1}^2. Similarly, a transitive relation for the error terms of the second scale is simultaneously produced. Such an offset by one graduation mark produces only an additional n equations, and the rank of the resulting system is $2n$ with $2n + 1$ unknowns. In fact, no permutation possible by rigid body transformations on the two scales gets rid of this rank deficiency, which must be resolved by the introduction of some external information, i.e. scale information. Redundancy can be developed in the system of equations by sequentially introducing offsets of origin from the first to the mth graduation point in the calibrating range. Further, additional permutations can be obtained by 'scale reversal'—reversing a scale and aligning the nth point of one with the origin of the other. With the introduction of the scaling information, the transitivity obtained through the offset or reversal operation and the redundancy obtained as a result of repeated offset operations, a robust estimate of the errors can be obtained through a least-squares solution of the set of equations.

3. Self-calibration with dual actuation

Dual actuation of an axis is often used to extend its capabilities (e.g., resolution, range, bandwidth for small displacements). In this section, we adapt the basic idea of self-calibration for implementation on dual actuation. For nanopositioning systems, since sensors such as capacitive sensors are used instead of scales, we will refer to the measuring devices as 'sensors' instead of scales. As previously mentioned, each actuator's sensor plays the role of the external scale/artefact to the other, thus allowing for the self-calibration of the system. Figure 5 shows a schematic of the two drives and feedback sensors. The sensor configuration chosen is well adapted to typical flexure-based nanopositioning systems. In such systems, because of the typically small displacements involved, both sensors can be connected to the inertial frame. In other situations, different sensor configurations may be used, but the general approach does not change. In this study, it is assumed that the first sensor (sensor 1) plays the role of the feedback sensor for the lower stage and the second sensor (sensor 2) works with the upper, fine motion stage that is mounted on the lower stage's table. Because this second sensor is mounted in the inertial frame (i.e. the machine's base), it reads that the displacement of both stages' sensors is fixed to the ground. Conversely, a motion of the lower stage is recorded on both sensors, but the motion of the upper stage is only recorded on the second sensor.

To conduct self-calibration of the system, the two actuators are first moved so that both sensors register zero. This is taken as the origin of both actuation stages. The bottom stage is then moved through discrete uniform steps (corresponding to the graduation marks in the previous section), and the displacements are recorded on both sensors. The difference in the readings on the two sensors produces the values that correspond to the observations O_i in (3). After stepping through n such equally spaced positions, the lower stage is returned to its origin. The upper stage is now displaced through one step and the process is repeated with the lower stage taking n steps with the differences at the sensors being recorded to produce observations O_i^1 of (4). The process is repeated with the offset of the upper stage increasing progressively up to m steps. Each such repetition produces an observation set, O_i^j, of (4) with j increasing up to a value m.

This sequence of actuations and observations produces a set of equations that can be assembled into matrix form, $AE = O$, where A is a relational matrix of dimension [nm, $2n + m$], E is a vector of dimension [$2n + m$] consisting of errors at n points and $n + m$ equally spaced points on sensors 1 and 2, respectively, and O is a vector of dimension [nm], containing the observed differences in the readings of the sensors for each permutation of displacements at the two actuators. This assembly of equations into matrix form is as follows:

$$AE = O$$

$$\begin{matrix} \\ nm \end{matrix} \overbrace{\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & -1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 & -1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 & 0 & \cdots & -1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 & 1 & -1 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 1 & 0 & -1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 & 1 & 0 & 0 & -1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 1 & 0 & 0 & 0 & \cdots & -1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & 0 & -1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & \cdots & 0 & 1 & 0 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 & \cdots & 0 & 1 & 0 & 0 & 0 & \cdots & -1 \end{bmatrix}}^{n \qquad\qquad n+m}$$

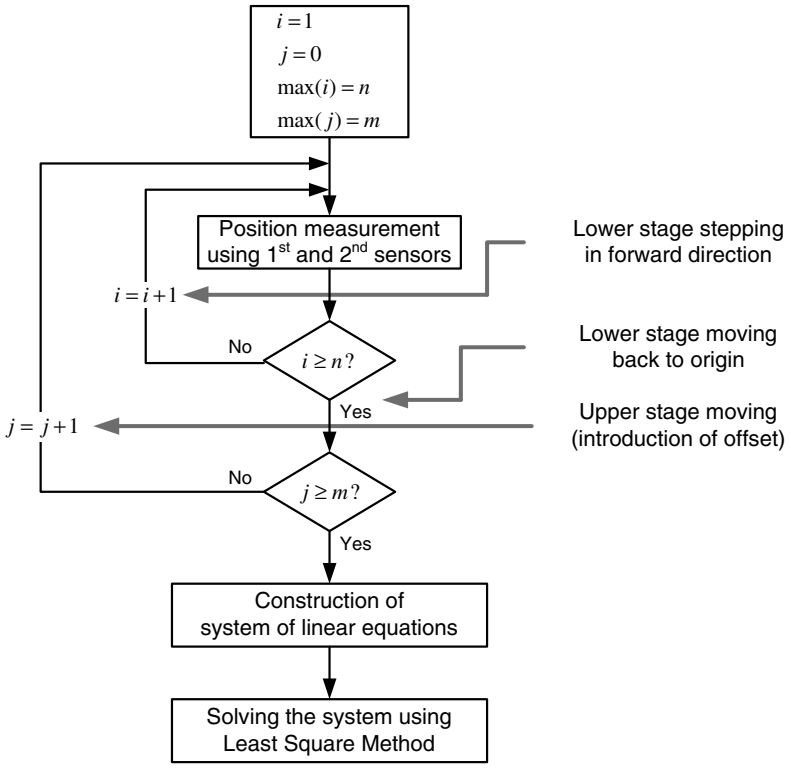


Figure 6. Flow chart of one-dimensional self-calibration for a dual stage.

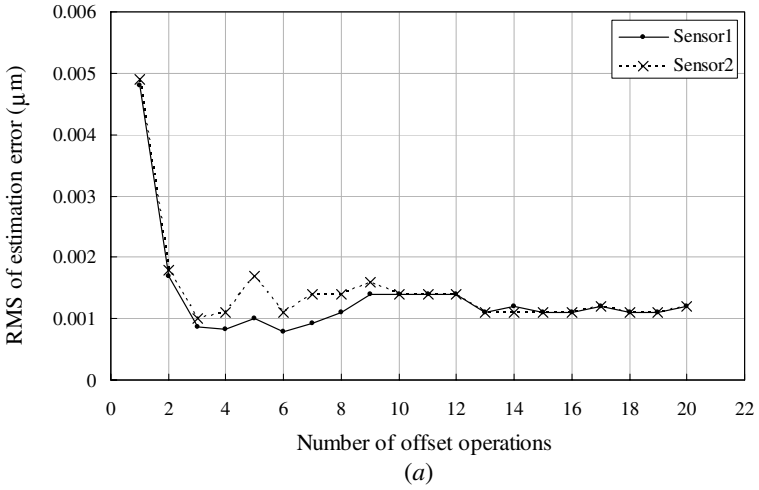
$$\times \begin{bmatrix} e_1^1 \\ e_2^1 \\ e_3^1 \\ \vdots \\ e_n^1 \\ - \\ e_1^2 \\ e_2^2 \\ e_3^2 \\ \vdots \\ e_n^2 \\ e_{n+1}^2 \\ e_{n+2}^2 \\ \vdots \\ e_{n+m}^2 \end{bmatrix} = \begin{bmatrix} o_1 \\ o_2 \\ o_3 \\ \vdots \\ o_n \\ - \\ o_1^1 \\ o_2^1 \\ o_3^1 \\ \vdots \\ o_n^1 \\ - \\ \vdots \\ - \\ o_1^m \\ o_2^m \\ o_3^m \\ \\ o_n^m \end{bmatrix} \tag{5}$$

As previously mentioned, only one offsetting (of the upper sensor relative to the lower sensor) operation is required to generate enough equations to solve for all unknowns (when some scale-defining information is input into the system). However to obtain some robustness against measurement errors, multiple offsets provide the requisite data redundancy. Irrespective of the number of offsets used, an analysis of the relational matrix, A, in (5) quickly establishes the need for the introduction of scale-defining information because the matrix remains rank deficient with a rank of $2n + m - 1$. Suppose such information is inserted in the form of knowledge of the error of the first graduation mark of sensor 1, then when that value is inserted for e_1^1, the result system in $2n + m - 1$ unknowns can be easily solved for. Alternatively, an additional equation containing this information can be introduced into the system of (5). When this system has redundant equations (because of multiple offsetting operations), it can be solved using least squares, producing a solution given by

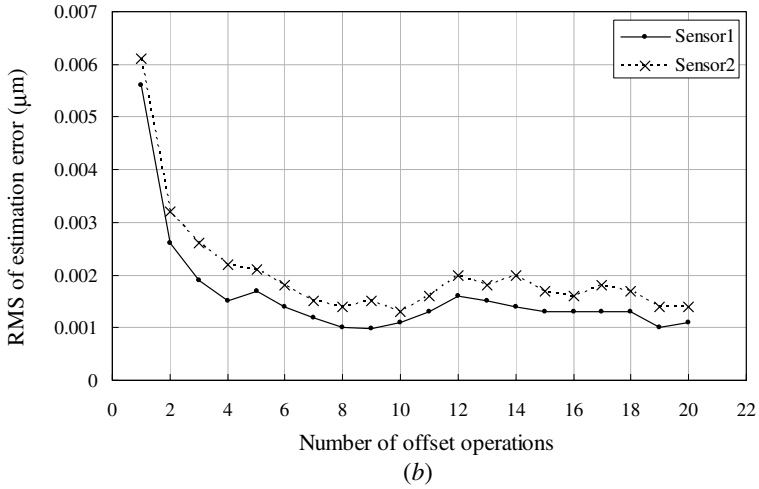
$$E = (A^T A)^{-1} A^T O. \tag{6}$$

When solving the system using least squares, care must be given to how the scaling information is inserted. If it is inserted as an additional equation into the system of (5), it will only be given a weight equal to every other equation in the system and therefore not be enforced exactly (i.e. the equation may have a non-zero residual in the least-squares solution). If exact enforcement is desired, the set of (5) must be reduced to a system in $2n + m - 1$ unknowns before solving it by the least-squares procedure.

In this way, self-calibration can be implemented on a dual-actuated axis without the use of an external artefact or scale but, instead, by using the internal actuation and sensing redundancy of the axis itself and only one external piece of information for scaling to a traceable length standard. As mentioned earlier, this information may be input into the system by means of a simple artefact with one feature/dimension calibrated



(*a*)



(*b*)

Figure 7. The variation of RMS errors with the number of offset operations: (*a*) when the noise has a mean of 0 and standard deviations of 1.5 nm for sensor 1 and 1.0 nm for sensor 2, (*b*) when the noise has a mean of 0 and standard deviations of 1.0 nm for sensor 1 and 3.0 nm for sensor 2.

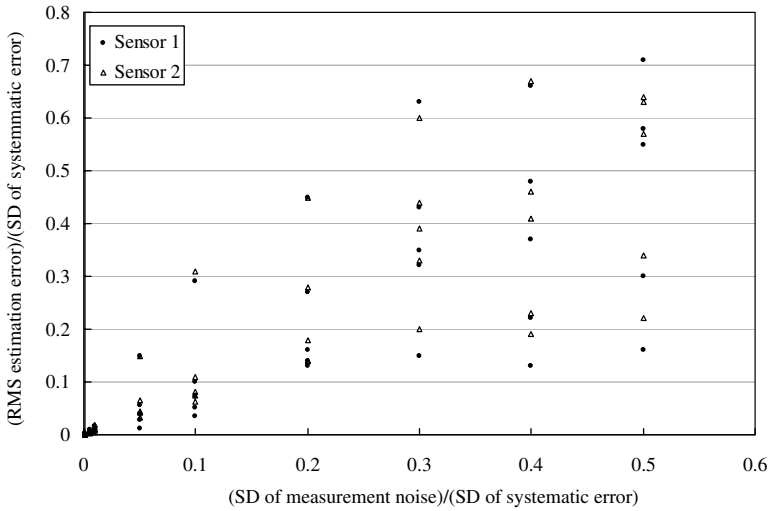


Figure 8. The variation of the ratio of the RMS value of estimation errors to the standard deviation of the introduced systematic error with the ratio of standard deviation of measurement noise to the standard deviation of introduced systematic error. The introduced systematic errors are randomly drawn for a normal distribution with a zero mean and a standard deviation of 0.01 μm. No averaging of measurements is performed.

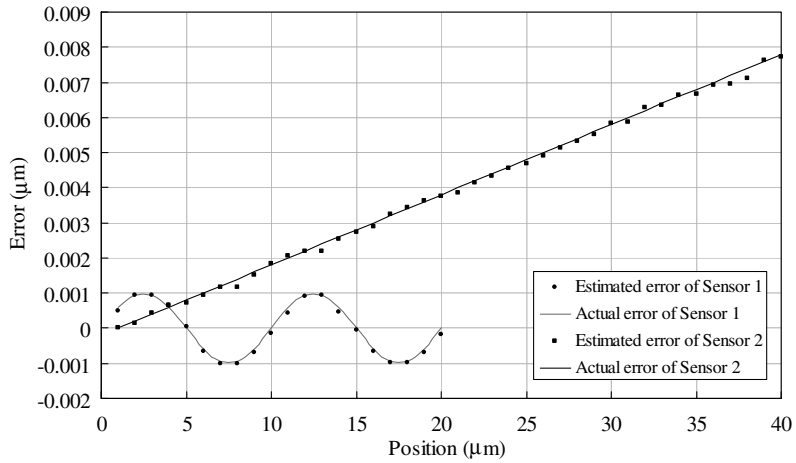


Figure 9. Simulation result when sensor 1 has sinusoidal and sensor 2 has linear systematic error characteristics.

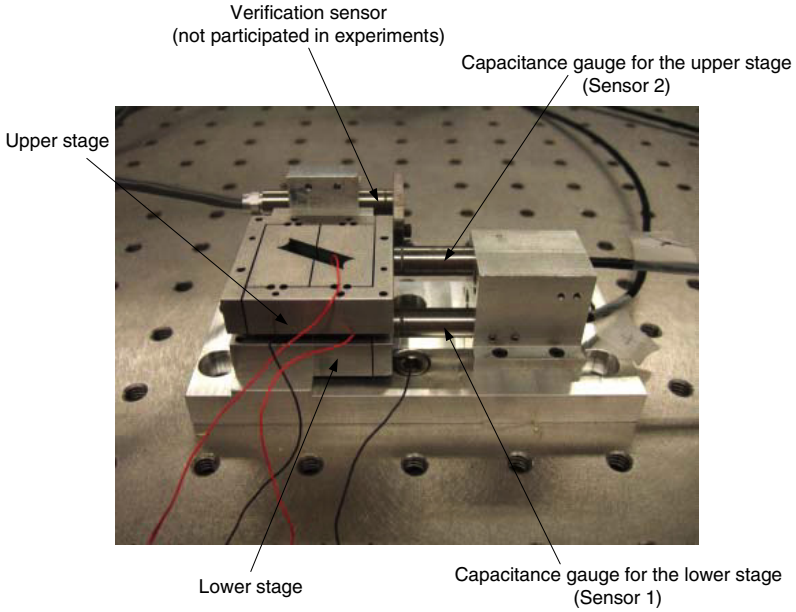


Figure 10. Experimental set-up of a dual-stage positioning system with two capacitance sensors.

to a traceable length standard. The flowchart in figure 6 outlines the procedure for self-calibration of a dual-actuated axis.

4. Simulation and experimental verification

The behaviour of the approach proposed in section 3 for self-calibration of dual-actuated nanopositioning stages is studied using simulations for its ability to deal with random measurement errors and different systematic error characteristics. The physical scheme modelled is the same as that shown in figure 5 and discussed in the previous section. In the simulation, it is assumed that the sensors produce analogue outputs which contain a random component modelled by a normal distribution with the mean of 0 and a specified standard deviation.

To study the effectiveness of increasing the measurement redundancy by increasing the number of offset steps (note that each offset step on the upper actuator introduces 'n' new equations, where n is the number of steps carried out by the lower actuator) and to determine typically how many offset operations are needed, two simulations are performed with different levels of measurement noise and repeated for an increasing number of offsets. In these simulations, the number of steps at which the lower sensor is calibrated, n, is 20 and that for the upper sensor, $n + m$, is 40 and the step size is 1 μm. The systematic errors are drawn from the functions given in equations (7) and (8) before the simulation of error measurement at each of the $2n + m$ points on the two sensors. Whenever a reading of a sensor at a particular point is simulated during the execution of the self-calibration simulation, this pre-determined value is added to the nominal

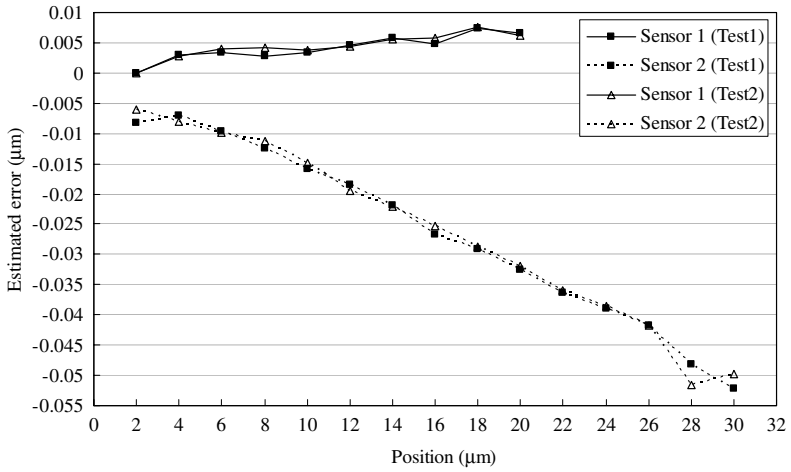


Figure 11. Calibration results for two sensors of a dual-stage positioning system

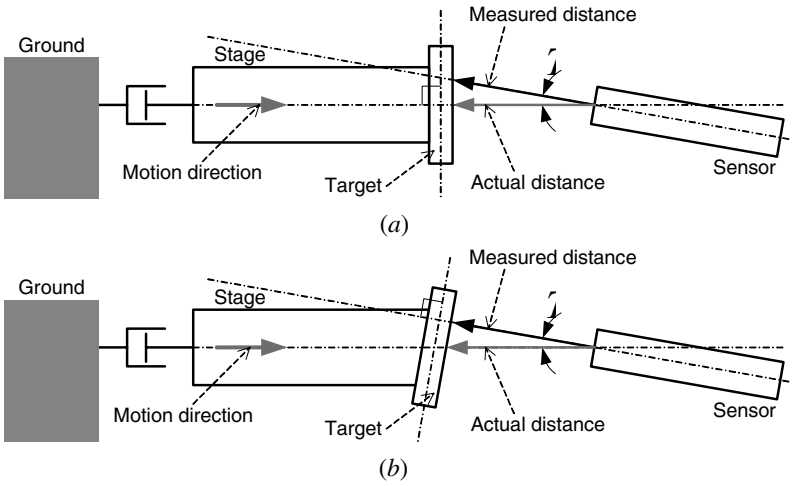


Figure 12. Influence of misalignments between the sensor, target and actuator: (*a*) when only the sensor is misaligned, (*b*) when the sensor and the target are misaligned.

displacement associated with the point. In addition, during each simulated reading of a sensor a simulated noise term is also added. The observations are simulated as per the steps outlined in the flowchart of figure 6. Figure 7(*a*) shows the RMS (root mean square) value of the estimation error of the self-calibration process for different numbers of offsets when the systematic errors were drawn from (7) and (8) and the measurement noises at sensors 1 and 2 were $N(0, 1.5)$ and $N(0, 1.0)$, respectively, where the standard deviation is in nanometres. Figure 7(*b*) shows the same results when the noise levels at sensors 1 and 2 are $N(0, 1.0)$ and $N(0, 3.0)$ respectively. For the case when the offset is 1, the system of equations is solved exactly. For all other offsets, least squares is used to estimate the sensor errors. From the results, it is clear that the self-calibration procedure with more than three offset steps performs at a level slightly better than the measurement noise level with very slow improvement (probably due to the averaging effect of increased redundancy) with number of offsets.

To understand the susceptibility of the approach to measurement noise, known systematic errors drawn from a $N(0, 10.0)$ distribution were introduced into the system. Seven simulations (with $n = m = 20$) were then conducted, each with an increasing level of measurement noise. The first experiment had no noise. In the remaining six, the RMS of measurement noise is increased relative to that of the systematic error by drawing the measurement noise from a distribution $N(0, \mathrm{SD})$, with SD increasing gradually from 0.5 to 5 nm (or from 5 to 50% of the standard deviation of the distribution that generated the systematic error). Five repetitions of the simulation experiment were conducted at each measurement noise level. Figure 8 shows how the RMS value of estimation error normalized against standard deviation of the distribution from which the systematic errors were drawn varies as the ratio of the standard deviations of the measurement noise to systematic errors. The fact that the ratio stays below 1 for noise levels approaching 50% of the signal (systematic errors) suggests a robust calibration technique.

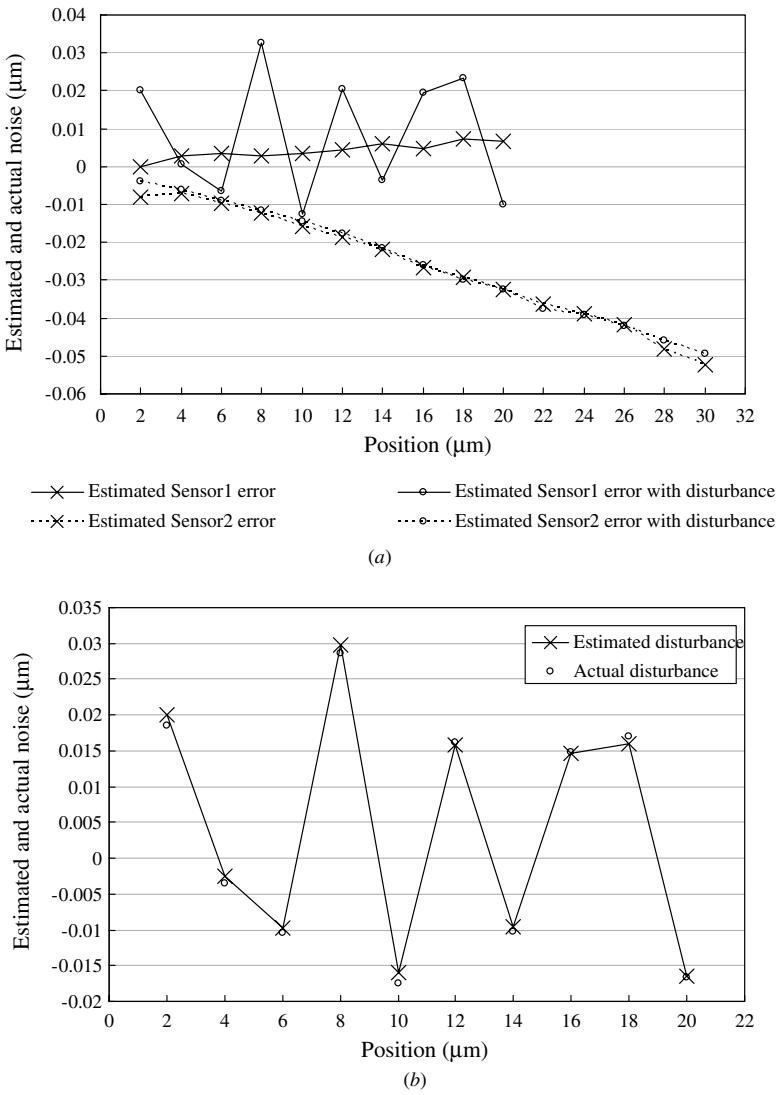


Figure 13. Verification of the original estimated systematic error using a known disturbance. (*a*) Estimated error with the disturbance overlaid on the original estimated errors. (*b*) Comparison between the introduced and estimated disturbances when the original estimated errors are subtracted.

To demonstrate the estimation performance for known functional forms of the systematic sensor error, a simulation is performed. In the simulation, the calibration interval was set to be 1 μm with the calibration ranges of 20 μm for sensor 1 and 40 μm for sensor 2. The measurement noise is modelled as a random signal normally distributed with the mean of 0 and the standard deviation of 2 nm. The data are averaged with 100 samples at every calibration point to simulate a typical measurement that would be conducted to take care of the high-frequency measurement noise. The simulation uses a sinusoidal systematic error for sensor 1 and a linear error characteristic for sensor 2, as follows:

$$e_i^1 = 0.001 \cdot \sin\left(\frac{\pi}{5}\left(x_i^1\right)\right) \tag{7}$$

$$e_i^2 = 0.0002 \cdot x_i^2 - 0.0002. \tag{8}$$

Similar scaling information input is used as the first two simulations. The simulation results are shown in figure 9, in which the estimated error successfully tracks the modelled error with the maximum estimation error of 0.17 nm in sensor 2. The RMS of estimation errors of sensors 1 and 2 are 0.084 and 0.15 nm, respectively. From these simulation results, it may be concluded that the self-calibration method for dual-stage positioning is capable of resolving the errors of the two sensors with high fidelity, with the RMS of the estimation error being less than 11% of the measurement noise.

With the simulations indicating robust and consistent behaviour for the self-calibration method with three or more offset steps, an experimental verification of the method on a dual-actuated nanopositioner was attempted. Figure 10 shows a photograph of the experimental set-up. The dual-actuated nanopositioner is realized by stacking two modular piezo-driven parallelogram 4-bar single-degree-of-freedom flexure stages [23] together. Each stage has a maximum travel range of 40 μm. The sensor configuration has two capacitance sensors

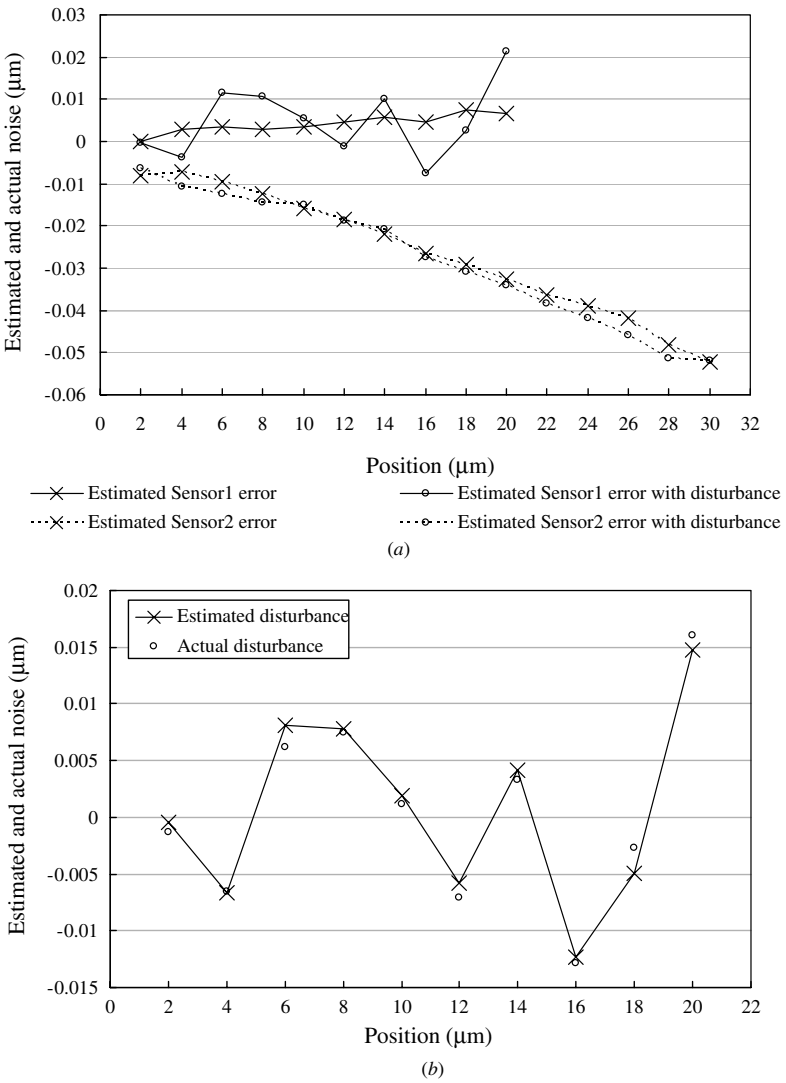


Figure 14. Verification of the original estimated systematic error using a known disturbance. (*a*) Estimated error with the disturbance overlaid on the original estimated errors. (*b*) Comparison between the introduced and estimated disturbances when the original estimated errors are subtracted.

that are to be calibrated. They are mounted in the base frame in exactly the same manner as depicted in the schematic of figure 5. A Delta-Tau UMAC Turbo controller is used for the control of the system. The two capacitance gauges (ADE Corporation, Gauge type: 4810) have a linearity of about 11.54 and 11.48 nm, an accuracy of 17.00 and 13.09 nm and a noise level of 0.87 and 0.89 nm, respectively. The UMAC Turbo controller drives the piezo actuators (APC Internationals Ltd, Pst150/5×5/7) using an amplifier (CEDRAT technologies, LC 75A) with the amplification ratio of 20. In figure 10, the additional sensor seen at the side of the set-up is used to verify the initial zero position for the two stages; it does not participate in the process of self-calibration, but rather it is there for the convenience of running the calibration experiment.

Calibration ranges for the upper (sensor 2) and lower (sensor 1) stages/sensors are 30 and 20 μm, respectively. Each calibration point or graduation mark spacing is set to 2 μm. Therefore, the numbers of calibration points for lower and upper stages/sensors are 10 and 15. The high-frequency sensor noise was around $\pm$20 nm. Therefore, at every calibration point the position was measured using the two capacitance gauges with the sampling frequency of about 73 Hz for 55 s. The obtained values were averaged to minimize the influence of a high-frequency sensor noise. To check the repeatability of the calibrated results, the experiments were performed twice on two different days. Figure 11 shows the estimated errors of upper and lower sensors. The maximum error of the lower sensor is 6.3 nm at 20 μm away from its origin, while the upper sensor has the maximum error of $-$50.4 nm at 30 μm away from its origin. The errors of both sensors seem to vary almost linearly with displacement but with slopes of different signs. The cases of positive and negative slopes may possibly be generated by assembly errors, as shown in figure 12. The error of the lower sensor has a positive slope suggesting that the sensor is inclined at an

angle of 0.0251 radians relative to the direction of motion, assuming that the target is correctly mounted. In the case of the lower sensor which has a negative slope error, the target and sensor have an inclination of 0.0580 radians with respect to the direction of motion, assuming that the target and sensor are correctly aligned with each other. The difference between the estimation results obtained from two repetitions of the self-calibration procedure is 1.3 nm (lower sensor) and 3.4 nm (upper sensor). The RMS of the differences are 0.62 nm (lower sensor) and 1.38 nm (upper sensor). Therefore, one may conclude from these observations that the self-calibration procedure produces very repeatable calibration results.

The question that arises from such a self-calibration process is whether the estimated errors are artefacts of the experiment and the self-calibration procedure or whether they are really present in the system. To some extent, the consistency of the results in the two repetitions of the experiment suggests the latter. To further verify their existence, a third experiment was conducted in which an intentional known perturbation of errors was superimposed on to the system (at sensor 1), as in [22]. These perturbations are drawn randomly from a normal distribution with the mean of 0 and the standard deviation of 20 nm, a value between the absolute maximum values of the errors of the two sensors. This known perturbation of error is introduced into each command of the lower stage so that

$$X^1_{\text{command},i} = X^1_{\text{nominal},i} + \Delta X^1_i, \tag{9}$$

where $X^1_{\text{command},i}$, $X^1_{\text{nominal},i}$ and ΔX^1_i are the actual command to a lower stage, nominal command position and the known disturbance given intentionally at the ith calibration point when the system of linear equations is constructed. While the values of ΔX^1_i are known to the experimenter, they are not known to the self-calibration process and the estimation algorithms it uses. Therefore, they should show up as superimposed over the true error characteristics of sensor 1. Other than this, the calibration process is repeated in precisely the same manner as the previously described experiment. Figure 13(*a*) shows the errors identified on sensor 1 by the self-calibration procedure after the known perturbations were introduced along with the original error characteristics of the sensor (the average of the errors identified in the two previous repetitions of the self-calibration process). Figure 13(*b*) shows how the identified perturbations match up to the introduced perturbations, after the originally identified errors of the sensor are subtracted out. Figure 14 shows similar results for a second repetition of the experiment with different values of perturbations introduced to the systematic error, drawn from the same distribution as before. The close match—the RMS error is 1.2 nm and the maximum difference is about 2.2 nm—and the repeatability of the results give us confidence both that the self-calibration process works and that the error that is identified is in fact present on the sensor and is not an artefact of the identification process.

5. Conclusions

This paper has addressed how the external artefact-based single-axis self-calibration technique can be adapted to self-calibration of a dual-actuated single axis without the need for an external artefact that explicitly defines each measurement point (the external scaling information input to a traceable length standard is always required, and this may be introduced into the system by a simple artefact). The approach exploits the redundancy in sensing and actuation of such systems to provide the transitive relationships needed to pass the external scale information to the different calibration points in the system. Needing no external artefact and hence no positioning and repositioning, the approach can easily be automated and run as a procedure in the start-up sequence of an automated tool. The approach developed is particularly applicable in the context of nanopositioning technology where obtaining a stable artefact with a well-defined feature and positioning and probing it with high repeatability can be challenging and time consuming.

The self-calibration process developed in this paper has been shown to be effective through both simulations and experiments. The procedure allows for the introduction of any desired number of redundant observations through the selection of appropriate procedure parameters (in this case, the number of offsetting steps), allowing the user to tune the procedure's ability to alleviate the effects of measurement noise. Using this parameter, the process was shown to be capable of reducing its RMS of the estimation error to a level below the standard deviation of the measurement noise. In experiments, the approach was able to reproducibly identify error characteristics to within a fraction of a nanometre, using relatively standard commercial off-the-shelf measurement and actuation components. A very high level of repeatability was observed in different sets of experiments conducted at different times, in spite of a relatively high level of measurement noise and only nominal control of environmental disturbances. This repeatability was found to be around 1 nm (RMS error) over three different experimental runs, indicating a robust self-calibration procedure. Besides the applicability of this self-calibration procedure for nanopositioning stages, it is also applicable to conventional stages and positioning systems. Future work will address multi-axis (XY and XYZ) stages, in which the interaction between different axes, such as squareness and perpendicularity errors, contributes to the systematic errors of the system.

Acknowledgments

This material is based upon work supported by the National Science Foundation through the Center for Nanoscale Chemical Electrical and Mechanical Manufacturing Systems under Award Number DMI 0328162 and through Award Number Grant DMI 0422687. Financial assistance was also obtained through Technology Research, Education and Commercialization Center, funded by the Office of Naval Research and administered by the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign. Dr Jeong was supported in part by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2006-214-D00007)

References

[1] Raugh M R 1996 Overlay can be improved by self-calibrated XY measuring instrument: a lattice perspective *Proc. SPIE* **2884** 379–91

[2] Phillips S D, Estler W T, Doiron T, Eberhardt K R and Levenson M S 2001 A careful consideration of the calibration concept *J. Res. Natl Inst. Stand. Technol.* **106** 371–9

[3] 2004 *International Vocabulary of Basic and General Terms in Metrology* 3rd edn (Geneva: International Organization of Standards)

[4] Hocken R J and the Machine-Tool Task Force 1980 *Technology of Machine-Tools* vol 5 *Machine Tool Accuracy* UCRL-52960-5 (Livermore, CA: Lawrence Livermore Laboratory)

[5] Kiridena V S B and Ferreira P M 1994 Kinematic modeling of quasistatic errors of three axis machining centers *Int. J. Mach. Tools Manuf.* **34** 85–100

[6] Donmez M A, Blomquist D S, Hocken R J, Liu C R and Barash M M 1986 A general methodology for machine tool accuracy enhancement by error compensation *Precis. Eng.* **8** 187–96

[7] Ferreira P M and Liu C R 1986 A contribution to the analysis compensation of the geometric error of a machining center *Ann. CIRP* **35** 259–62

[8] ISO 230-1 1996 *ISO 230-1: Geometric Accuracy of Machines Operating Under No-Load or Finishing Conditions* (Geneva: International Organization of Standards)

[9] ISO 230-2 2006 *ISO 230-2: Determination of Accuracy and Repeatability of Positioning Numerically Controlled Axes* (Geneva: International Organization of Standards)

[10] Hocken R *et al* 1993 *ANSI/ASME B5.54: Methods for Performance Evaluation of Computer Numerically Controlled Machining Centers* (New York: American National Standards Institute)

[11] Beers J S and Penzes W B 1999 The NIST length scale interferometer *J. Res. Natl Inst. Stand. Technol.* **104** 225–52

[12] Jørgensen J F, Jensen C P and Garnaes J 1998 Lateral metrology using scanning probe microscopes, 2D pitch standards and image processing *Appl. Phys.* A **66** 847–52

[13] Markiewicz P and Goha M C 1995 Atomic force microscope tip deconvolution using calibration arrays *Rev. Sci. Instrum.* **66** 3186–90

[14] Bogdanov A L 2007 Metrology challenges in nanofabrication of photonics devices *Tri-National Workshop on Standards for Nanotechnology (Ottawa)*

[15] Evans C J, Hocken R J and Estler W T 1996 Self-calibration: reversal, redundancy, error separation, and 'absolute testing' *Ann. CIRP* **45** 617–34

[16] Raugh M R 1997 Two-dimensional stage-calibration: role of symmetry and invariant sets of points *J. Vac. Sci. Technol.* B **15** 2139–45

[17] Raugh M R 1997 Error estimation for lattice methods of stage self-calibration *Proc. SPIE* **3050** 614–25

[18] Raugh M R 1985 Absolute two-dimensional sub-micron metrology for electron beam lithography *Precis. Eng.* **7** 3–13

[19] Raugh M R 1986 Auto-calibration method suitable for use in electron beam lithography *US Patent* 4583298

[20] Takac M 1993 Self-calibration in one-dimension *SPIE Photomask Technol. Manage.* **2087** 80–6

[21] Lee M C and Ferreira P M 2002 Auto-triangulation and auto-trilateration: I. Fundamentals *Precis. Eng.* **26** 237–49

[22] Lee M C and Ferreira P M 2002 Auto-triangulation and auto-trilateration: II. Three-dimensional experimental verification *Precis. Eng.* **26** 250–62

[23] Dong J, Jeong Y H and Ferreira P M 2007 Design and analysis of a micro-positioning module for multi-degree-of-freedom micro-positioners *Int. Conf. on Micromanufacturing (Greenville, SC)*

IOP PUBLISHING JOURNAL OF MICROMECHANICS AND MICROENGINEERING

J. Micromech. Microeng. **18** (2008) 035011 (10pp) doi:10.1088/0960-1317/18/3/035011

Simultaneous actuation and displacement sensing for electrostatic drives

Jingyan Dong and Placid M Ferreira

Department of Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, 1206 W Green Street, Urbana, IL 61801, USA

E-mail: pferreir@uiuc.edu

Received 16 September 2007, in final form 12 December 2007
Published 24 January 2008
Online at stacks.iop.org/JMM/18/035011

Abstract

This paper presents a method for driving a MEMS electrostatic actuator, while simultaneously sensing the resulting displacement/capacitance without the use of an additional physical sensing structure. The approach superposes the sensing and actuation signals into a single input into the system and obtains its mechanical (displacement) response from the modulation (amplitude or phase) it produces on the sensing input. The approach is analyzed and experimentally shown to produce an amplitude modulation of 0.1857 mV μm^{-1} of displacement on electrostatic drive that produces a displacement of 14 μm at 100 V and a 0.55 pF capacitance change from a nominal capacitance of 0.35 Pico farads. The approach enables a very cost-effective and convenient approach to detect the displacement of MEMS devices for a variety of applications in the laboratory environment, and provide a potential feedback signal for closed-loop control of electrostatically driven MEMS devices.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

In MEMS (micro electro mechanical system) devices, capacitive structures have been widely used in both actuation (as electrostatic or comb drives) and sensing applications. Electrostatic actuators (see figure 1), including parallel plate actuators [1, 2], lateral comb drives [3–7] and vertical comb drives [8], provide a means of producing controlled translation or rotational displacements. In the context of silicon-based MEMS devices, unlike the other actuation technologies, electrostatic actuators avoid extra processing steps and the need for integrating additional materials, such as shape memory alloys, piezoelectric film/actuators or magnets/coils. Thus they have been used in many applications, such as variable optical attenuators (VOA) [1], micro/nano positioning stages [2–4], scanning probe microscopes [5, 7], data storage [6, 7] and optical mirrors [8].

For active MEMS devices, it is very important to monitor the displacement of the moving parts (manipulators). These displacements either reflect a measurement of certain physical/chemical properties or values when the device interacts with the environment, or provide precise positioning information for desired tasks. For the case that the motion is out of the plane for optical applications, optically based methods can conveniently be used to characterize the capabilities of a device. For example, optical interferometric surface profilometers [8] are used to get static measurements of vertical motion and laser Doppler vibrometers [1, 8] are used for dynamic measurements of out-of-plane motions. The interferometry-based displacement detection approaches generally have high measurement resolution, but are limited by their response speed because of the need for noise filters and averaging. Further, because of the size of the components used, they are difficult to use in the measurement of lateral displacements for MEMS applications. For in-plane comb drives, microscopy is often used to calibrate displacement [3–5]. While conventional optical microscopes are limited by the diffraction limit to about a 0.5 μm resolution, scanning probe microscopes such as near field scanning optical microscopes (NSOM) and atomic force microscopes (AFM) can be used for higher resolutions. Scanning electron microscopy (SEM) is not suitable for such measurements because of interaction between the electrostatic actuators and the electron beam. Calibration of lateral displacements is also done by calibration artifacts such as gratings and other repeating patterns [25]. In general, these techniques are good for calibrating displacements of MEMS devices, but cannot be

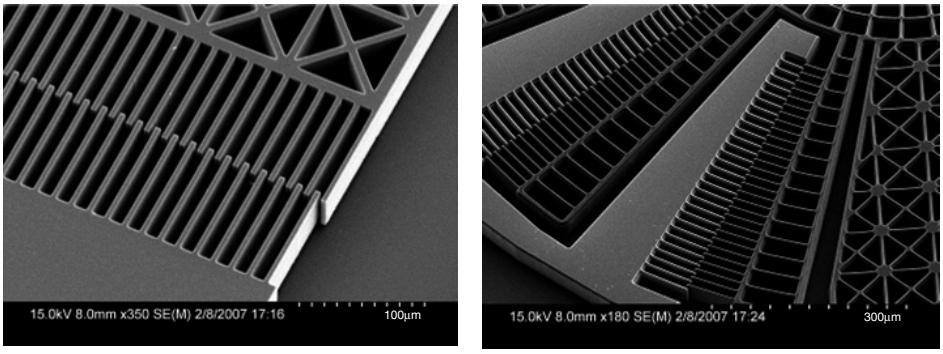


Figure 1. Examples of electrostatic or comb drives. Left: linear comb drive; right: rotary comb drive.

used for closed-loop control or measurements during operation and interaction with other structures.

Capacitive sensing provides a potentially high-precision, high-sensitivity displacement detection method. The capacitance of a structure changes as the relative position of two electrodes on the structure changes. This effect has been used widely in capacitive sensors, such as capacitive pressure sensors [9], capacitive bending strain sensors [10], accelerometers [11], position sensors [27] and encoders for long-range position sensing [12, 13]. Due to the scale of MEMS devices, the capacitance changes are generally very small ($<$1 pF). Parasitic capacitance from the measurement setup makes precise detection of tiny capacitance changes a challenging issue. Commercially available LCR capacitive meters are used to measure the capacitance from strain sensor [10] and positioning sensors [14, 15]. Resonant LCR circuits [9, 15, 22] are also used to detect the change of capacitance in comb drives. A resonant LCR circuit is built by connecting an inductor and a resistor to the capacitive structure, in this case a comb drive. As the capacitance changes, the resonant behavior changes accordingly; this change is used to detect the capacitance change. For the packaged MEMS devices, electronics circuits are integrated onto the same chip so as to minimize the parasitic capacitance and achieve high measurement integrity. An electronic readout interface converts the measured capacitance to a voltage for position readout. Charge amplifiers [12, 13, 21] or differential charge amplifiers [18, 19, 28] are used to measure capacitance by measuring the charge flow with respect to reference voltage from an oscillator. The generated dc voltage output is proportional to the charge flow going through the measured capacitor and therefore proportional to the measured capacitance. These circuit designs are optimized to fit for different requirement [16, 17]. Besides the above charge amplifier based circuits, ac bridge circuits [20] are also used to measure small capacitances. The difficulty lies in adjusting the bridge circuit to balance both magnitude and phase.

In most previous research, actuation and capacitive sensing are implemented through separate physical structures. In such schemes, there is an actuation structure, such as an electrostatic drive [2, 12, 13, 22, 24, 26] or thermal actuator [15], as well as a separate sensing capacitive structure, most typically, a comb structure. For MEMS devices, the sensing structure increases the overall size of the device and takes up precious real-estate decreasing the favorable economics of parallel fabrication of typical lithography-based micromachining processes. Further, it adds to the moving mass of the system, resulting in a decrease of the resonant frequency and hence the frequency range of operation of the device. Also, the additional flexural structures such as hinges or leaf springs associated with the sensing structure add mechanical resistance that reduces the displacement range of the device. It would therefore be advantageous to use the same capacitive structure, such as a single comb drive, simultaneously for both actuation and sensing purposes.

This paper proposes, develops and experimentally verifies a new approach to simultaneously actuate a MEMS electrostatic drive while at the same time sensing its resulting displacement. Section 2 presents the concept behind and analysis of the approach. Section 3 presents experiments on the combs of a stage to demonstrate the functioning of the approach. Section 4 draws up conclusions and remarks about applications.

2. Scheme for simultaneous sensing and actuation of an electrostatic drive

The sensing scheme developed here exploits two important aspects of such electrostatic drives: the large difference in frequency ranges in which the drive structure is mechanically and electrically responsive and the dependence of drive capacitance on its mechanical state (displacement). Mechanically, the drive behaves as a second-order mass–spring–damper system, attenuating mechanical response to a negligible level for input signals whose frequencies are orders of magnitude higher than the mechanical resonant frequency. This makes it possible to superimpose such a high-frequency sensing signal on the actuation signal. Electrically, the drive is an RC circuit with a variable capacitance, which changes with mechanical displacement of the drive. This sensing signal, while evoking no mechanical response, experiences a significant amplitude and phase modulation due to the capacitance change of the drive (due, in turn, to its mechanical displacement in response to the actuation signal). By monitoring these changes across the resistor or capacitor of the circuit, relative to the input signal, the capacitance change

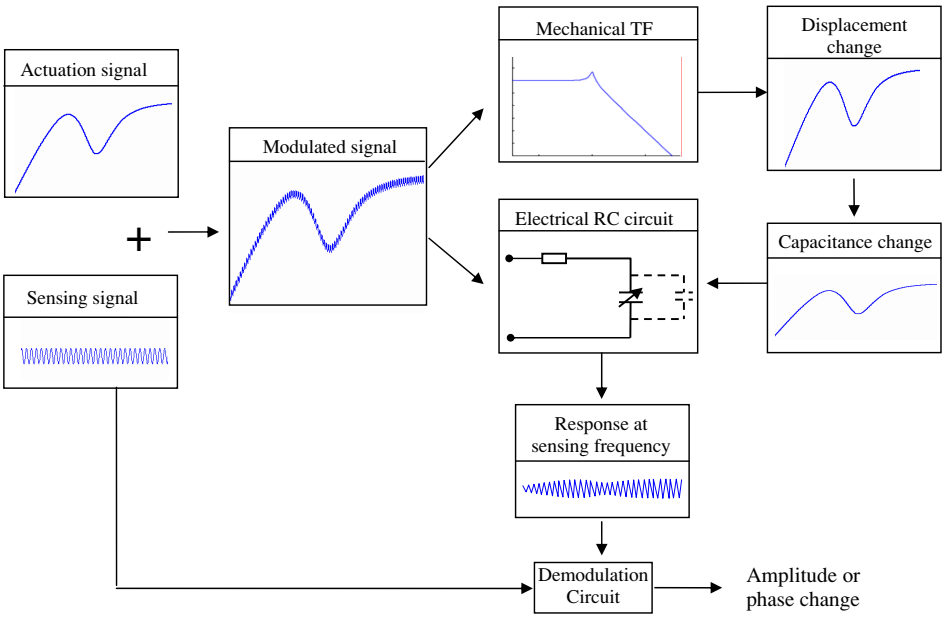


Figure 2. Displacement/capacitance measurement scheme based on amplitude modulation and amplitude/phase detection.

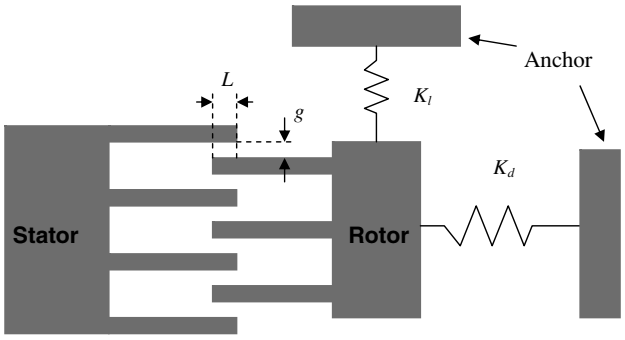


Figure 3. Typical comb actuator with geometrical parameters.

and hence displacement of the drive can be obtained through a demodulation circuit. Thus, no additional sensing structure is required by this strategy. Figure 2 schematically shows the approach to simultaneous actuating and displacement sensing for a capacitive drive.

Throughout this paper, comb drives are analyzed as representative of electrostatic actuators. Figure 1 shows two typical comb drive structures, while figure 3 shows the parameters associated with such drives. The interdigitated fingers between the fixed stator and the movable rotor form the capacitive structure. The rotor is connected to anchors by flexible suspending structures, such as folded springs. The stiffness of the suspending structures is represented by a spring constant K_d in the actuation direction and a spring constant K_l in the lateral direction. Generally, K_l is much larger than K_d so as to avoid side instability.

For a comb actuator with n fingers in the rotor (assuming no fringe effects), the overall force is $F = n\frac{\varepsilon t}{g}V^2$, where ε is the permittivity of free space, t is the thickness of interdigitated fingers, g is the gap between fingers and V is the supply voltage.

Assuming that the stiffness along the actuation direction is constant, K_d, which is generally valid when the actuated displacement is much less than the compliant dimension of the suspension structure, the static displacement of the actuator is $d = \Delta L = \frac{F}{k_d}$. This displacement causes the capacitance in the gap of the structure to change by $\Delta C = \frac{\varepsilon t}{g}\Delta L = \frac{\varepsilon t}{g}d$, suggesting a proportional relationship between the state (displacement) of the drive and its capacitance. Further, during actuation, in addition to the capacitive driving force and the restoring spring force of the suspension, the comb drives must overcome the initial force of the mass of the moving structure, the viscous and squeeze damping forces due to the air in the interdigitated structures and other interactions the structure may have with the environment [2]. Therefore the drive's mechanical response is governed by a second-order mass–spring–damping dynamic system as follows:

$$m\ddot{d} + f\dot{d} + K_d d = F \qquad \text{or}$$
$$\frac{d(s)}{F(s)} = \frac{1}{ms^2 + fs + K_d} = \frac{k}{s^2 + 2\xi\omega_n s + \omega_n^2} \tag{1}$$

where m is the equivalent mass of the device, f is the damping coefficient, $k = 1/m$, $\omega_n = \sqrt{K_d/m}$ is the mechanical natural or resonant frequency, $\xi = \frac{f}{2\sqrt{mK_d}}$ is the damping ratio. The normalized (with the horizontal axis set to the ratio $\frac{\omega}{\omega_n}$) magnitude bode plot for such a second-order system, shown in figure 4, indicates that when the frequency of the input (voltage, for this comb drive system), $\omega > \omega_n$, the gain or magnitude of the transfer function (in this case the ratio of amplitude of mechanical displacement to input voltage) decays at a rate of 40 dB (1/100) per decade. Thus for an input signal with a frequency 100 times larger than the resonant frequency ω_n, the magnitude of the response is 10^{-4} times that of a dc or low frequency input with the same amplitude. This difference in mechanical response to the inputs with different

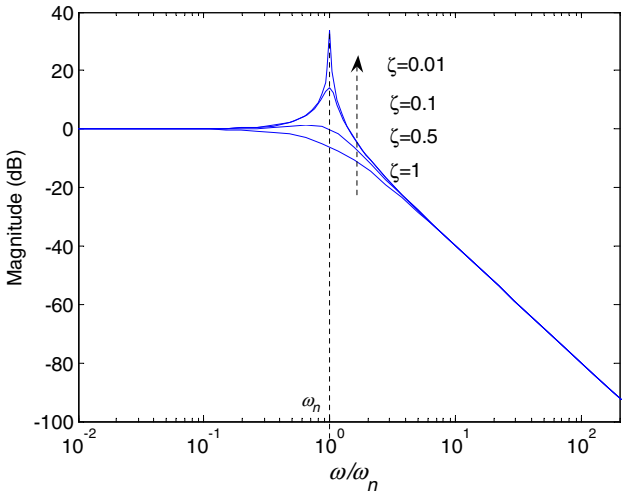


Figure 4. Normalized magnitude Bode plot of the comb actuator.

frequencies, in particular the severe attenuation of mechanical displacements (and hence change in capacitance of the comb) in response to high frequency signals, enables the possibility of simultaneously actuating and probing the mechanical response (sensing the displacement) of an electrostatic structure.

Having established that the high frequency signal evokes negligible mechanical response from the drive, we now analyze the nature of the sensing signal, its magnitude and the magnitude of modulation one can expect (due to changes in the electrical parameters of the circuit representing the drive as a result of mechanical displacement) to assess the feasibility of such a simultaneous actuation and sensing strategy. Figure 5 shows a schematic of a circuit that implements this approach. A voltage adder circuit is built using a high voltage operational amplifier. The output from the operational amplifier is easily verified to be $V_{\text{out}} = \bar{V}_{\text{Act}} + \tilde{V}_{\text{Sen}}$, the addition of the (low frequency) high voltage actuation signal and the high-frequency, constant (low) voltage sensing signal. A resistor R_{load} is connected in parallel with the RC circuit as a load to protect the electrostatic drive by providing current/charge surges with a bypassed path instead of through the drive. The other resistor R_1 is connected in series with the comb drive and forms a RC circuit with the comb drive providing the capacitance. The value of R_1 is chosen to be comparable to the impedance of the capacitor at the sensing frequency.

The actuation signal is a low-frequency signal (for many applications that involve positioning this is a dc signal, for others that involve scanning as in SPMs, in the range of 0–100 Hz). At low frequencies, due to its small capacitance, the comb drive has very high impedance. Now if the measurement or sensing is done at a frequency of 100 kHz, the impedance of the comb drive in the actuation frequency band is about 1000 times than that at the sensing frequency. Since the resistor across which the measurements are made is chosen to have a value of impedance close to that of the comb drive at the sensing frequency, its impedance is very small compared to that of the drive in the actuation frequency band (about 1000 times smaller). Thus most of the voltage drop in the actuation frequency band occurs across the drive and a very small drop occurs across the resistor. The magnitude of voltage drop on the resistor as a result of the actuation voltage is $V_{R_1} = \left|\frac{R_1}{1/j\omega C_c + R_1}\right| V_{\text{Act}} = \frac{\omega R_1 C_c}{\sqrt{1+(\omega R_1 C_c)^2}} V_{\text{Act}}$. For a MEMS-scale capacitive drive system, C is in the range of 0.5 pF. If the resistor R_1 is 1 MΩ; this turns out to be $V_{R_1} = 0.0003\, V_{\text{Act}}$ at 100 Hz and $V_{R_1} = 3 \times 10^{-6}\, V_{\text{Act}}$ at 1 Hz. For a maximum actuating input of about 100 V, the voltage drop on the resistor is less than 0.03 V. Thus the sensing circuit is subjected to a very small fraction of the actuation input. An important consequence is that, should an on-chip implementation of the sensing circuit be undertaken, conventional low voltage designs such as those given in [29–31] can be implemented.

When actuated, the capacitance of the comb structure changes accordingly with $\Delta C = \frac{\varepsilon t}{g}\Delta L = \frac{\varepsilon t}{g} d$. At the sensing frequency, the impedance due to the comb drive capacitance, $1/j\omega C$, is smaller than in the actuation frequency band and comparable to that of the resistance. The change of capacitance of the comb drive due to mechanical displacement therefore produces a more pronounced change in the behavior of the RC circuit at this frequency, introducing an amplitude and phase change on the voltage drop across the resistor

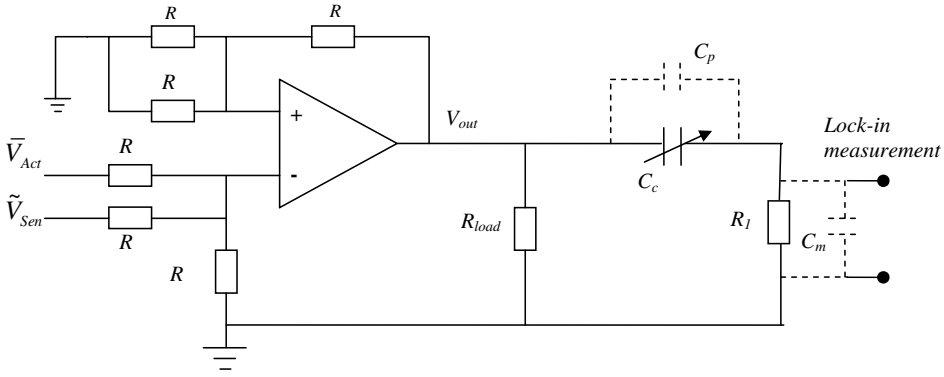


Figure 5. Schematic circuit for *in situ* displacement sensing with actuation.

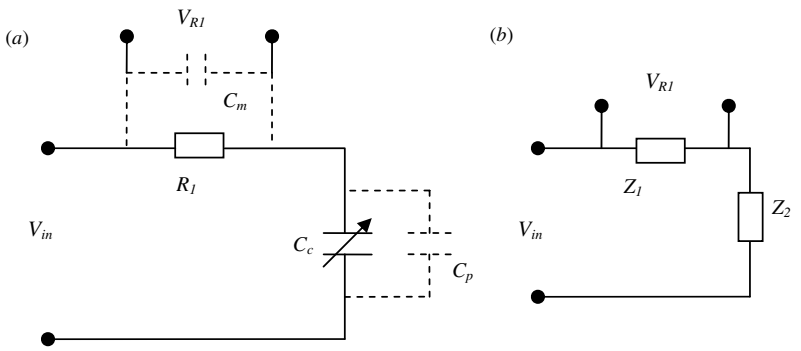


Figure 6. A RC circuit with parasitic capacitances (dashed line) (*a*) and its equivalent simplified circuit (*b*).

or capacitor at the sensing frequency relative to the sensing reference (or input) signal. Since the reference frequency is a parameter that is selected, this amplitude or phase change can be picked up by a lock-in circuit with very high precision as an indication of displacement or capacitance change.

Under ideal conditions, when parasitic capacitance is negligible, both the phase or amplitude change over the resistor R_1 can be used to detect the change of the capacitance. Referring to figures 6,

$$V_{R_1} = \frac{Z_1}{Z_1 + Z_2} V_{\text{in}} = \frac{\text{j}\omega R_1 C_c}{1 + \text{j}\omega R_1 C_c} V_{\text{in}},$$

thus

$$G_{R_1} = \left| \frac{V_{R_1}}{V_{\text{in}}} \right| = \frac{\omega R_1 C_c}{\sqrt{1 + (\omega R_1 C_c)^2}} \tag{2}$$

$$\phi_{R_1} = \tan^{-1} \frac{1}{\omega R_1 C_c}. \tag{3}$$

The maximum sensitivity (signal variation) is obtained when $Z_1 = Z_2$, that is $R_1 = \frac{1}{\omega C_c}$.

Unfortunately, in the laboratory and typical operational environments, parasitic capacitance is not negligible when compared with the capacitance from comb drives, typically in the range of 0.1–1 pF. A coaxial BNC cable normally brings in 1 pF cm^{-1} parasitic capacitance and a lock-in amplifier has more than 20 pF in input capacitance. Additionally, the probe station, commonly used a test setup in labs, can also contribute parasitic capacitance from its cable. These parasitic capacitance sources are introduced in the measurement system as C_m and C_p, as shown in figure 6. With the parasitic capacitance in the system, the voltage across the resistor can be expressed by equation (4),

$$\begin{aligned} V_{R_1} &= \frac{Z_1}{Z_1 + Z_2} V_{\text{in}} = \frac{\frac{R_1}{1+\text{j}\omega R_1 C_m}}{\frac{R_1}{1+\text{j}\omega R_1 C_m} + \frac{1}{\text{j}\omega (C_p + C_c)}} V_{\text{in}} \\ &= \frac{\text{j}\omega R_1 (C_p + C_c)}{1 + \text{j}\omega R_1 (C_p + C_c + C_m)} V_{\text{in}} \end{aligned} \tag{4}$$

$$G_{R_1} = \left| \frac{V_{R_1}}{V_{\text{in}}} \right| = \frac{\omega R_1 (C_p + C_c)}{\sqrt{1 + [\omega R_1 (C_p + C_c + C_m)]^2}} \quad \text{and}$$

$$\phi_{R_1} = \tan^{-1} \frac{1}{\omega R_1 (C_p + C_c + C_m)}, \tag{5}$$

where Z_1 is the impedance of the combined resistor R_1 and the parasitic capacitance C_m, Z_2 is the impedance of the combined comb capacitance C_c and parasitic capacitance C_p, ω is the sensing frequency.

If parasitic capacitance from the measurement circuit is very large when compared with that from the device, the phase change would be small but the amplitude change can still be significant. When $C_m \gg C_c$, corresponding to a certain displacement or capacitance change $\text{d}C_c$, the gain change or change in magnitude of the output signal with unit input signal is $\text{d}G_{R_1} \cong \frac{\omega R_1}{\sqrt{1+[\omega R_1 (C_p+C_c+C_m)]^2}} \text{d}C_c = k(\text{d}C_c)$. Thus the change of the amplitude is proportional to the change of capacitance.

Sensitivity is critically important for any measurement method. For this measurement method, the sensitivity is increased when the amplitude change of the voltage across the resistor is made as large as possible when compared with its initial amplitude. The amplitude change ratio can be expressed by equation (6),

$$\frac{\text{d}G_{R_1}}{G_{R_1}} \cong \frac{\omega R_1}{\omega R_1 (C_p + C_c)} \text{d}C_c = \frac{\text{d}C_c}{C_p + C_c}. \tag{6}$$

From equation (6), the smaller the parasitic capacitance C_c, the larger the amplitude change ratio, leading to a more precise measurement.

To detect the amplitude change, two schemes may be used, as shown in figure 7. The first scheme is based on a lock-in circuit. A lock-in circuit (including a reference circuit, a multiplier and a low-pass filter) tuned to the input sensing frequency can easily detect amplitude changes with high precision. Since such a monitoring strategy only picks up a signal at the reference frequency, little measurement noise enters the measurement signal, providing for a very precise measurement. The scheme is well suited to high precision, static position measurement even in the presence of large noise. The other scheme, better suited to low-noise situations, also allows for dynamic positioning and tracking control. Since the sensing signal has a known frequency, a band-pass filter centered at the signal's frequency followed by a low-pass filter will convert displacement changes into a dc voltage. The lock-in and filter circuits are both implementable on a chip using schemes such as those in [29–31] to provide high signal integrity and make the system a stand-along device

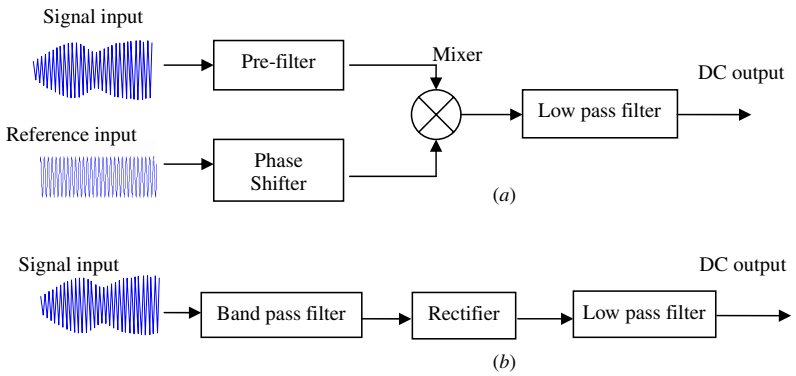


Figure 7. Two schemes for displacement signal recovery. (*a*) A basic simplifier lock-in circuit based displacement detection (for signals with large noise). (*b*) Direct filter based displacement detection (for clean signals with little noise).

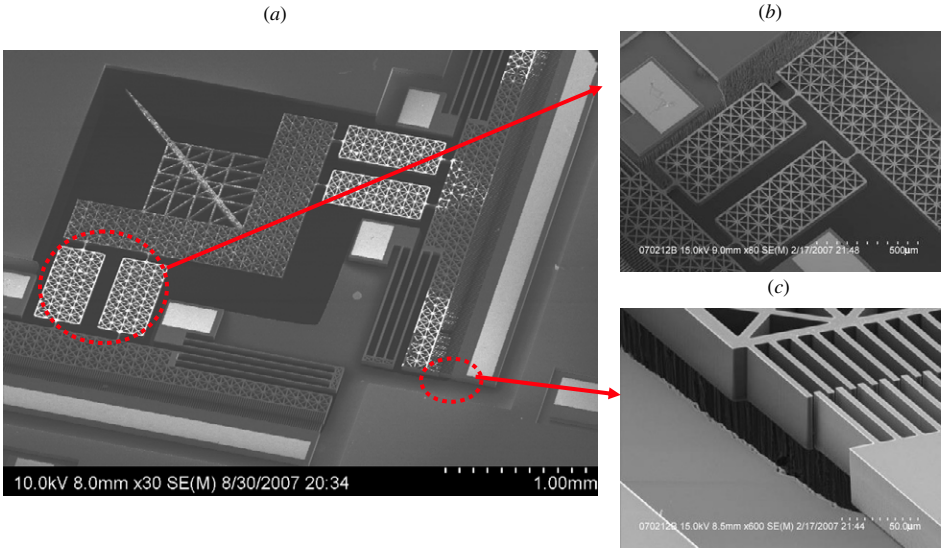


Figure 8. PKM micro positioning *XY* stage used for testing. (*a*) Overall structure; (*b*) four bar linkage mechanism with flexure hinges; (*c*) comb actuator and fingers.

with both actuation and sensing capabilities. In this paper, to demonstrate the scheme for simultaneous actuation and sensing on a single capacitive structure, a commercial lock-in amplifier is used.

3. Test setup and experimental results

The simultaneous actuation and sensing strategy for electrostatic drives is tested a SOI MEMS parallel kinematic *XY* micropositioning stage (figure 8) with an improved design, from what was reported in [4], to increase the motion range and suppress unwanted parasitic motions. In this parallel kinematic mechanism design, there are two independent kinematics chains that connect the end effector to the base (stator). Each of these kinematic chains includes two serially-connected joints, a controlled prismatic joint implemented by a linear comb diver, and a parallelogram 4-bar linkage joint, which maintains the orientation of the end effector invariant. Due to the parallel kinematic design, the stage has a high natural frequency (more than 1 K Hz) and a motion range about 14 μm at a driving voltage of 100 V. A probe is integrated into the *XY* stage as a functional manipulator. The targeted applications, such as materials (thin film) characterization and mechanical testing of biological structures, make precise position sensing very important. The fabrication of this device includes three patterning steps and two etching steps. The comb structures are fabricated by the DRIE Bosch process. The handle layer at the back of the device is also etched away so that the test sample can be feed from either the top or the bottom. The detailed fabrication steps are discussed in [4].

The fabricated comb drive has 220 pair fingers, thus about 440 gaps. The thickness of the fingers is 50 μm and the gap between fingers is about 5 μm. The initial engagement of the interdigitated fingers is about 9 μm. With the above parameters, the initial capacitance from the comb

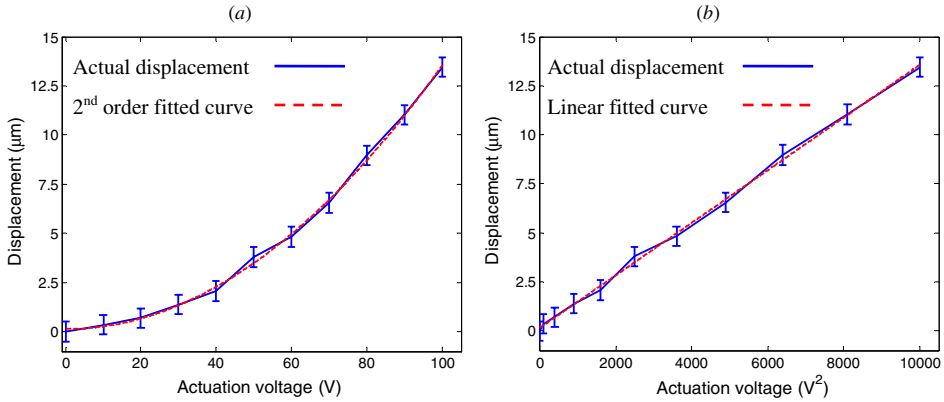


Figure 9. Displacement of the comb actuator observed from a microscope at different actuation voltages as a function of driving voltage (*a*) and driving voltage square (*b*).

drive is about 0.35 pF. If the maximum displacement of the comb drive is 15 μm, then the capacitance change will be only 0.58 pF. Compared with the parasitic capacitance from the measurement loop including coaxial cables and lock-in amplifier, which is generally several hundreds pF, the capacitance change of comb drive is extremely small. This large parasitic capacitance in parallel with the resistor from the measurement loop increases the difficulty for detecting the small capacitance change, such as the phase detection method. As discussed in the previous section, when there is a large parasitic capacitance in parallel with the resistor, the amplitude change can be used as an indicator of capacitance change. However, in order to increase the sensitivity, the parasitic capacitance in parallel with the comb drive has to be minimized. In our experiment, a probe station is used to connect the signal to the electrodes of the comb actuator. To decrease the parasitic capacitance from the wire connections, single-wire cables are used instead of coaxial or tri-axial cables. The cables are routed far away from each other and their lengths are minimized to what is essential. In this way, the parasitic capacitance in parallel with the comb drive is controlled to be less than 10 pF.

The sensing signal is obtained as a reference signal from the lock-in amplifier (SR850 from Stanford Research Systems, Inc) with a frequency of 100 KHz, which is about 100 times large than the natural frequency of the device. The sensing signal has a magnitude of 1 V; thus it induces a mechanical vibration amplitude is only 1/10 000 of 1 V dc input. Since the device moves approximately 14 μm at 100 V, the 1 V, 100 kHz sensing signal only moves the mechanism by negligible amount of 1.4×10^{-7} μm. The actuation voltage is supplied by a voltage amplifier (Trek 623 B) and commanded by a function generator (HP/Agilent 33220A). All the circuits are implemented on a breadboard and connected with the lock-in amplifier, power supply and probe station.

The driving voltage for an actuator is gradually incremented and the corresponding displacements of a comb drive are observed visually by tracking the motion of a feature on the end-effector with a microscope scale that has a resolution of 1 μm. At the same time, the amplitude of the voltage across the resistor of the RC circuit is measured by the lock-in amplifier at the sensing frequency. Figure 9 shows the static displacement of the stage at different voltages, as observed through the microscope. The experimental data overlay the second order fitted parabola curves for displacement. Within the resolution of the microscope, the displacement follows a parabolic relationship with the driving voltage (figure 9(*a*)) or a linear relationship with the square of the driving voltage (figure 9(*b*)). Due to the limitation of the optical microscope, the resolution of the displacement is low, as shown by the error bars in figure 9.

Next, the same experiment of increasing the actuation voltage in steps is carried out with the amplitude of the output from the lock-in amplifier recorded. The amplitude change from the lock-in amplifier is shown in figure 10. The actual reading fits a second-order parabolic relationship with the driving voltage V with a very high degree of conformity (figure 10(*a*)), or a linear relationship to V^2 (figure 10(*a*)). By offsetting the output and adjusting the sensitivity of the lock-in amplifier, a maximum amplitude change about 2.6 mV with a resolution of 0.1 μV is achieved. Since 2.6 mV amplitude change corresponding to about 14 μm displacement, a theoretical sub-nanometer (0.54 nm) displacement sensing resolution can be claimed for this experiment. Further, the system has a measurement gain of 0.1857 mV μm^{-1}.

Figure 11 demonstrates the ability of the approach in tracking a sinusoidal actuation input. A 1 Hz sinusoidal command is generated by a function generator and a resulting actuating voltage with amplitude of 40V is supplied to the comb actuator with the sensing signal (figure 11(*a*)), the measurement of the amplitude change is recorded and plotted in figure 11(*b*). The amplitude of the tracking signal matches well with that predicted from the static step measurements shown in figure 10. Clearly the sensing signal reflects the actuating effect correctly. In our scheme, since the sensing frequency is much higher than the actuation frequency, the time constant of the sensing response is much shorter than the time constant of the actuator response, thus the sensing method will give correct position measurement even when the drive capacitance keeps changing during the transition.

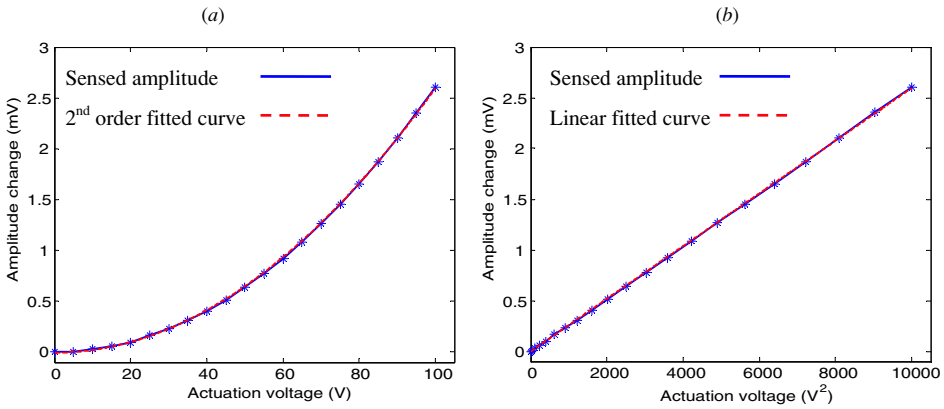


Figure 10. Amplitude change from the resistor observed the lock-in amplifier at different actuation voltages as a function of driving voltage (*a*) and driving voltage square (*b*).

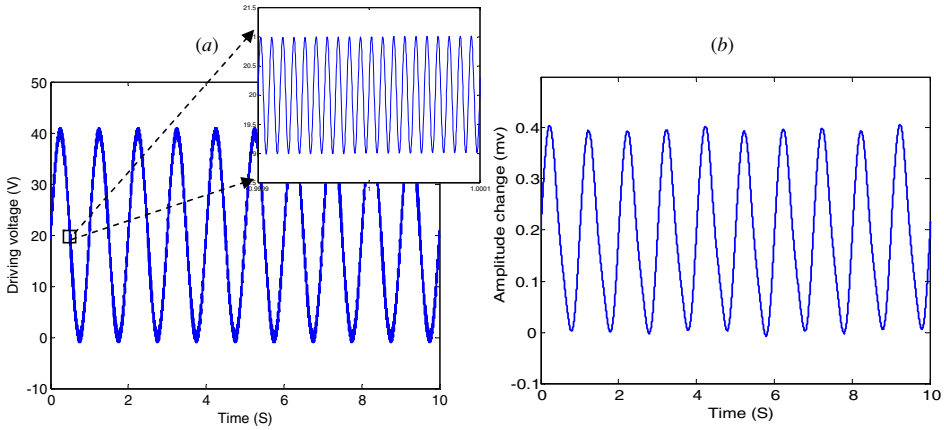


Figure 11. Amplitude change (*b*) observed by the lock-in amplifier for a 1 Hz sinusoid actuation voltage superposed with a sensing signal (*a*).

4. Conclusions

In this paper, a new approach to *in situ* simultaneous displacement sensing and actuation of electrostatic drives is presented. This method provides an efficient and powerful scheme to actuate the electrostatic actuators and sense the resulting displacement at the same time on a single comb structure. The sensing scheme is based on amplitude modulation of a high frequency sensing signal produced when the capacitance of the drive changes because of mechanical displacement. It exploits the fact that, mechanically, the electrostatic actuator behaves as a second-order (as a mass-spring-damping system) low-pass filter. This provides a high frequency region where the mechanical effects of a probing or sensing signal has no effect on the mechanical state of the actuator, but is capable of encoding it. A RC circuit is built with the electrostatic actuator as a variable capacitor. A high frequency sensing signal (about 100 times larger than the natural frequency of the mechanical system) is superposed to on to the actuation signal (normally in a frequency band less than the natural frequency of the mechanical system) using a operational amplifier-based voltage adding circuit. The sensing circuit does not need to work at high voltages. The reason is that the high voltage is resident in the low frequency, actuation band of electrostatic actuator. The sensing signal is a low voltage, high frequency signal. In the low frequency band, the electrostatic actuator has very large impedance, causing a very high voltage drop to a negligible level across the physical capacitive structure. The simultaneous actuation and sensing scheme is experimentally verified for an actual comb-drive of a parallel-kinematics MEMS *XY* positioning stage with a mechanical natural frequency of about 1 kHz. The displacement of the comb modulated a 1 V 100 KHz sensing signal to produce a measurement gain of 0.1857 mV μm^{-1}. The advantage of this method is that it does not require an additional physical structure for sensing. Instead, it has potential for high resolution (of the order of a nanometer) displacement sensing by only the use of some additional external electronics such as an amplifier superimposes the actuation and sensing voltage and lock-in type sensing circuit that demodulates the system's mechanical response (using phase or amplitude) from the high-

frequency reference signal. Future work includes improving sensing speeds and developing closed loop control systems using the position change detected here as the feedback signal.

Acknowledgments

This material is based upon work supported by the National Science Foundation through the Center for Nanoscale Chemical Electrical and Mechanical Manufacturing Systems under Award Number DMI 0328162 and through Award Number Grant DMI 0422687. Financial assistance was also obtained through the Technology Research, Education and Commercialization Center, funded by the Office of Naval Research and administered by the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign.

References

[1] Isamoto K K, Morosawa K, Changho Chong A, Fujita H and Toshiyoshi H 2004 A 5-V operated MEMS variable optical attenuator by SOI bulk micromachining *IEEE J. Sel. Top. Quantum Electron.* **10** 570–8

[2] Sun Yu, Piyabongkarn D, Sezen A, Nelson B J and Rajamani R 2002 A high-aspect-ratio two-axis electrostatic microactuator with extended travel range *Sensors Actuators* A **102** 49–60

[3] Tang W C, Nguyen T-C H and Howe R T 1989 Laterally driven polysilicon resonant microstructures *Sensors Actutors* **20** 25–32

[4] Dong J, Mukhopadhyay D and Ferreira P M 2007 Design, fabrication and testing of silicon-on-insulator (SOI) MEMS parallel kinematics XY stage *J. Micromech. Microeng.* **17** 1154–61

[5] Indermuhle P-F, Jaecklin V P, Brugger J, Linder C, de Rooij N F and Binggeli M 1995 AFM imaging with an xy-micropositioner with integrated tip *Sensors Actuators* A **47** 562–5

[6] Lu Y, Pang C K, Chen J, Zhu H, Yang J P, Mou J Q, Guo G X, Chen B M and Lee T H 2005 Design, fabrication and control of a micro X-Y stage with large ultra-thin film recoding media platform *Proc. 2005 IEEE/ASME Int. Conf. on Advanced Intelligent Mechatronics: AIM 2005* pp 19–24

[7] Kim C-H, Jeong H-M, Jeon J-U and Kim Y-K 2003 Silicon micro XY-stage with a large area shuttle and no-etching holes for SPM-based data storage *J. Microelectromech. Syst.* **12** 470–8

[8] Dooyoung H, Huang S T-Y, Jui-Che Tsai, Toshiyoshi H and Wu M C 2004 Low-voltage, large-scan angle MEMS analog micromirror arrays with hidden vertical comb-drive actuators *IEEE J. Microelectromech. Syst.* **13** 279–89

[9] Babbitt K E, Fuller L and Keller B 1997 A surface micromachined capacitive pressure sensor for biomedical applications *Proc. 12th Biennial University/Government/ Industry Microelectronics Symp.* pp 150–3

[10] Aebersold J, Walsh K, Crain M, Voor M, Martin M, Hnat W, Lin J, Jackson D and Naber J 2006 Design, modeling, fabrication and testing of a MEMS capacitive bending strain sensor *J. Phys.: Conf. Ser.* **34** 124–9

[11] Chau K H-L, Lewis S R, Zhao Y, Howe R T, Bart S F and Marcheselli R G 1995 An integrated force-balanced capacitive accelerometer for low-G applications *8th Int. Conf. on Solid-State Sensors and Actuators 1995 and Eurosensors IX: Transducers '95* vol 1 pp 593–6

[12] Kuijpers A A, Krijnen G J M, Wiegerink R J, Lammerink T S J and Elwenspoek M 2006 A micromachined capacitive incremental position sensor: 2. Experimental assessment *J. Micromech. Microeng.* **16** S125–34

[13] Kuijpers A A, Wiegerink R J, Krijnen G J M, Lammerink T S J and Elwenspoek M 2004 Capacitive long-range position sensor for microactuators *17th IEEE Int. Conf. on Micro Electro Mechanical Systems* pp 544–7

[14] Lai Y, Bordatchev E V and Nikumb S K 2006 Metallic micro displacement capacitive sensor fabricated by laser micromachining technology *Microsyst. Technol.* **12** 778–85

[15] Chu L L and Gianchandani Y B 2003 A micromachined 2D positioner with electrothermal actuation and sub-nanometer capacitive sensing *J. Micromech. Microeng.* **13** 279–85

[16] Tavakoli M and Sarpeshkar R 2003 An offset-canceling low-noise lock-in architecture for capacitive sensing *IEEE J. Solid-State Circuits* **38** 244–53

[17] Lei S, Zorman C A and Garverick S L 2005 An oversampled capacitance-to-voltage converter IC with application to time-domain characterization of MEMS resonators *IEEE Sensors J.* **5** 1353–61

[18] Suster M, Jun G, Chaimanonart N, Ko W H and Young D J 2004 Low-noise CMOS integrated sensing electronics for capacitive MEMS strain sensors *Proc. IEEE Custom Integrated Circuits Conf. 2004* pp 693–6

[19] Garmire D, Choo H, Muller R S, Govindjee S and Demmel J 2006 MEMS process characterization with an on-chip device *Nanotech 2006: The Technical Proc. of the Nano Science and Technology Institute (Boston, MA, May)* vol 3, pp 550–3

[20] Pilla S, Hamida J A and Sullivan N S 1999 Very high sensitivity ac capacitance bridge for the dielectric study of molecular solids at low temperatures *Rev. Sci. Instrum.* **70** 4055–8

[21] K V Lazarov and E T Enikov 2006 Micro-mechatronics and MEMS: capacitive position detection, Microsystems Mechanical Design *Series: CISM International Centre for Mechanical Sciences* ed F DeBona and E T Enikov (Berlin: Springer) p 478

[22] van Spengen W M and Tjerk H Oosterkamp 2007 A sensitive electronic capacitance measurement system to measure the comb drive motion of surface micromachined MEMS devices *J. Micromech. Microeng.* **17** 828–34

[23] Tang W C, Lim M G and Howe R T 1992 Electrostatic comb drive levitation and control method *J. Microelectromech. Syst.* **1** 170–8

[24] Jeong H-M and Ha S K 2005 Dynamic analysis of a resonant comb-drive micro-actuator in linear and nonlinear regions *Sensors Actuators* A **125** 59–68

[25] Kim C-H and Kim Y-K 2002 Micro XY-stage using silicon on a glass substrate *J. Micromech. Microeng.* **12** 103–7

[26] Michael S-C Lu and Gary K Fedder 2004 Position control of parallel-plate microactuators for probe-based data storage *IEEE J. Microelectromech. Syst.* **13** 759–69

[27] Horenstein M N, Perreault J A and Bifano T G 2000 Differential capacitive position sensor for planar MEMS structures with vertical motion *Sensors Actuators* A **80** 53–61

[28] Lemkin M and Boser B E 1999 A three-axis micromachined accelerometer with a CMOS position-sense interface and digital offest-trim electronics *IEEE J. Solid-State Circuits* **34** 456–68

[29] Gnudi A, Colalongo L and Baccarani G 1999 Integrated lock–in amplifier for sensor applications *Proc. 25th European Solid-State Circuits Conf. 1999: ESSCIRC '99 (21–23 Sept.)* pp 58–61

[30] Ferri G, De Laurentiisa P, D'Amicob A and Di Nataleb C 2001 A low-voltage integrated CMOS analog lock-in amplifier prototype for LAPS applications *Sensors Actuators* A **92** 263–72

[31] Lu G N, Pittet P, Sou G, Carrillo G and Mourabit A E L 2003 A novel approach to implementing on-chip synchronous detection for CMOS optical detector systems *Analog Integr. Circuits Signal Process.* **37** 57–66

Harnessing biological motors to engineer systems for nanoscale transport and assembly

Living systems use biological nanomotors to build life's essential molecules—such as DNA and proteins—as well as to transport cargo inside cells with both spatial and temporal precision. Each motor is highly specialized and carries out a distinct function within the cell. Some have even evolved sophisticated mechanisms to ensure quality control during nanomanufacturing processes, whether to correct errors in biosynthesis or to detect and permit the repair of damaged transport highways. In general, these nanomotors consume chemical energy in order to undergo a series of shape changes that let them interact sequentially with other molecules. Here we review some of the many tasks that biomotors perform and analyse their underlying design principles from an engineering perspective. We also discuss experiments and strategies to integrate biomotors into synthetic environments for applications such as sensing, transport and assembly.

ANITA GOEL[1,2] AND VIOLA VOGEL[3]

[1]Nanobiosym Labs, 200 Boston Avenue, Suite 4700, Medford, Massachusetts 02155 USA; [2]Department of Physics, Harvard University, Massachusetts 02138, USA; [3]Department of Materials, ETH Zurich, Wolfgang Pauli Strasse 10. HCI F443, 8093 Zurich, Switzerland
e-mail: agoel@nanobiosym.com; viola.vogel@mat.ethz.ch

By considering how the biological machinery of our cells carries out many different functions with a high level of specificity, we can identify a number of engineering principles that can be used to harness these sophisticated molecular machines for applications outside their usual environments. Here we focus on two broad classes of nanomotors that burn chemical energy to move along linear tracks: assembly nanomotors and transport nanomotors.

SEQUENTIAL ASSEMBLY AND POLYMERIZATION

The molecular machinery found in our cells is responsible for the sequential assembly of complex biopolymers from their component building blocks (monomers): polymerases make DNA and RNA from nucleic acids, and ribosomes construct proteins from amino acids. These assembly nanomotors operate in conjunction with a master DNA or RNA template that defines the order in which individual building blocks must be incorporated into a new biopolymer. In addition to recognizing and binding the correct substrates (from a pool of many different ones), the motors must also catalyse the chemical reaction that joins them into a growing polymer chain. Moreover, both types of motors have evolved highly sophisticated mechanisms so that they are able not only to discriminate the correct monomers from the wrong ones, but also to detect and repair mistakes as they occur[1].

Molecular assembly machines or nanomotors (Fig. 1a) must effectively discriminate between substrate monomers that are structurally very similar. Polymerases must be able to distinguish between different nucleosides, and ribosomes need to recognize particular transfer-RNAs (t-RNAs) that carry a specific amino acid. These well-engineered biological nanomotors achieve this by pairing complementary Watson–Crick base pairs and comparing the geometrical fit of the monomers to their respective polymeric templates. This molecular discrimination makes use of the differential binding strengths of correctly matched and mismatched substrates, which is determined by the complementarity of the base-pairing between them.

Figure 1b illustrates the assembly process used by the DNA polymerase nanomotor. A template of single-stranded DNA binds to the nanomotor with angstrom-level precision, forming an open complex. The open complex can 'sample' the free nucleosides available. Binding of the correct nucleoside induces a conformational change in the nanomotor which then allows the new nucleoside to be added to the growing DNA strand[1]. The tight-fitting complementarity of shapes between the polymerase binding site and the properly paired base pair guarantees a 'geometric selection' for the correct nucleotide[2]. A similar mechanism is seen in *Escherichia coli* RNA polymerase, where the binding of an incorrect monomer inhibits the conformational change in the motor from an 'open' (inactive) to a 'closed' (active) conformation[3].

Ribosome motors carry out tasks much more complex than polymerases. Instead of the four nucleotide building blocks used by polymerases to assemble DNA or RNA, ribosomes must recognize and selectively arrange 20 amino acids to synthesize a protein. This fact alone increases the chance of errors. Nevertheless, ribosomes obviously work (and do so along the same principles of geometric fit

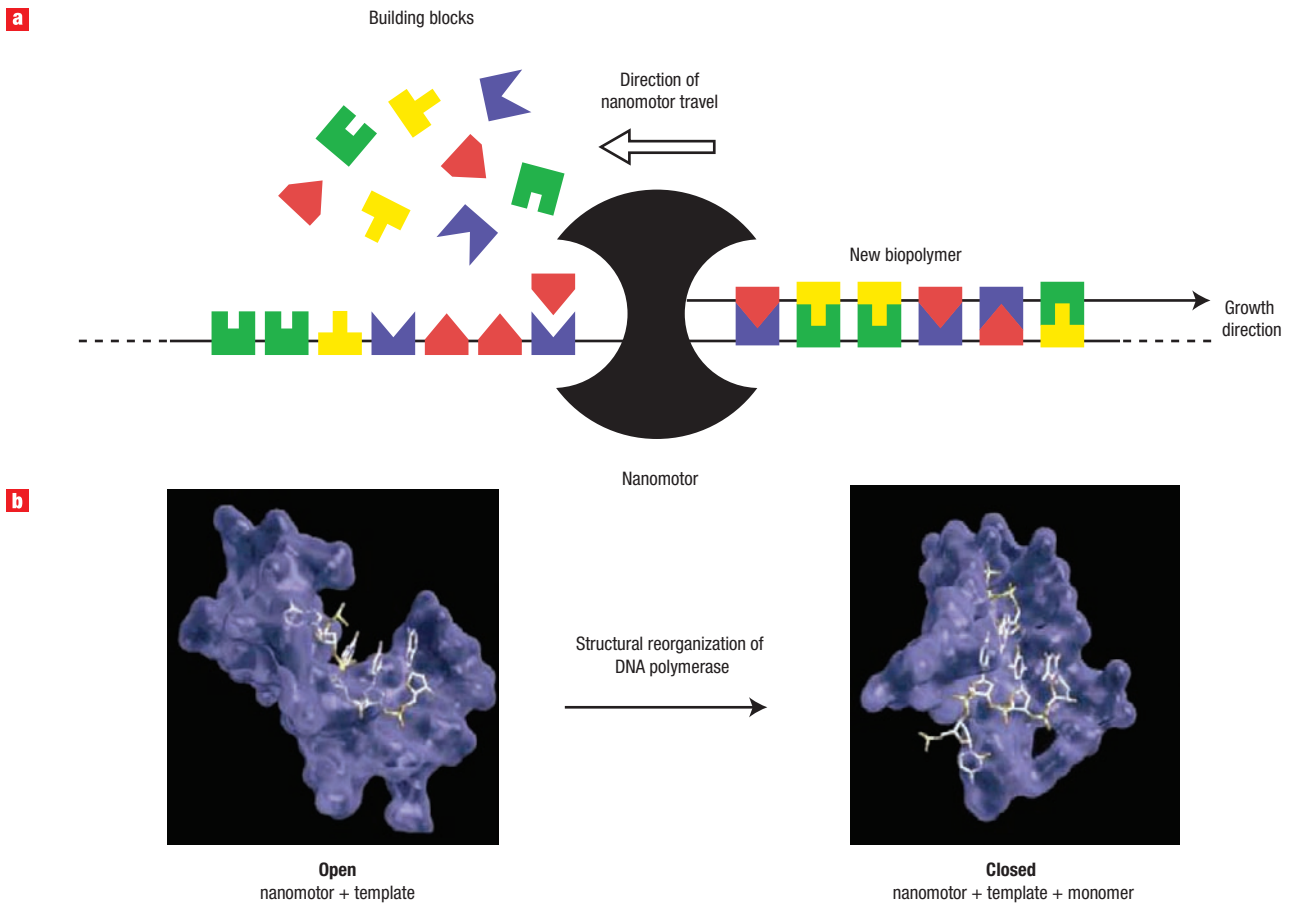


Figure 1 Molecular discrimination during sequential assembly. **a**, The polymerase nanomotor discriminates between four different building blocks as it assembles a DNA or RNA strand complementary to its template sequence. Molecular discrimination between substrate monomers that are structurally very similar is achieved by comparing the geometrical fit of the monomers to their respective polymeric templates. **b**, The T7 DNA polymerase motor undergoes an internal structural transition from an open state (when the active site samples different nucleotides) to a closed state (when the correct nucleotide is incorporated into the nascent DNA strand). Nucleotides are added to the nascent strand one at a time. This structural transition is the rate-limiting step in the replication cycle and is thought to be dependent on the mechanical tension in the template strand[2,9,107,116,121,127,128,131]. Figure adapted from ref. 127. Copyright (2001) PNAS.

and conformational change as do polymerases) and are able to build amino acid polymers that are subsequently folded into functional proteins. But ribosomal motors can be tricked, much more easily than DNA motors, into building the 'incorrect' sequences when supplied with synthetic amino acids that resemble real ones[4].

Engineering principle no. 1: Nanomotors used in the sequential assembly of biopolymers can discriminate efficiently between similar building blocks.

The structure of molecular machines can be visualized with angstrom-level resolution using X-ray crystallography, and the sequential assembly processes they drive can be probed in real time using single-molecule techniques[5–9]. By elucidating nanomotor kinetics under load, such nanoscale techniques provide detailed insights into the single-molecule dynamics of nanomotor-driven assembly processes. Techniques such as optical and magnetic tweezers, for example, have further elucidated the polymer properties of DNA[7,10–12]and the force-dependent kinetics of molecular motors[13–18]. Single-molecule fluorescence methods such as fluorescence energy transfer, in conjunction with such biomechanical tools, are illuminating the internal conformational dynamics of these nanomotors[19–21].

As the underlying design principles of assembly nanomotors are revealed, it will become increasingly possible to use these biomachines for *ex vivo* tasks. Sequencing and PCR are two such techniques that already harness polymerase nanomotors for the *ex vivo* replication of nucleic acids. The polymerase chain reaction, or PCR, is a landmark, Nobel prize-winning technique[22] invented in the 1980s that harnessed polymerase nanomotors to amplify a very small starting sample of DNA to billions of molecules. Likewise, there are many conceivable future applications that either use assembly nanomotors *ex vivo* or mimic some of their design principles. Efforts are already under way to control these nanomotors better and thus to improve such *ex vivo* sequential assembly processes for industrial use (see, for example, the websites www.cambrios.com; www.helicosbio.com; www.nanobiosym.com; www.pacificbiosciences.com).

In contrast, current *ex vivo* methods to synthesize block copolymers rely primarily on random collisions, resulting in a wide range of length distributions and much less control over the final sequence[23]. Sequential assembly without the use of nanomotors remains limited to the synthesis of comparatively short peptides, oligonucleotides and oligosaccharides[24–26]. Common synthesizers still lack both the precision of monomer selection and the inbuilt proofreading machinery for monomer repair that nanomotors have.

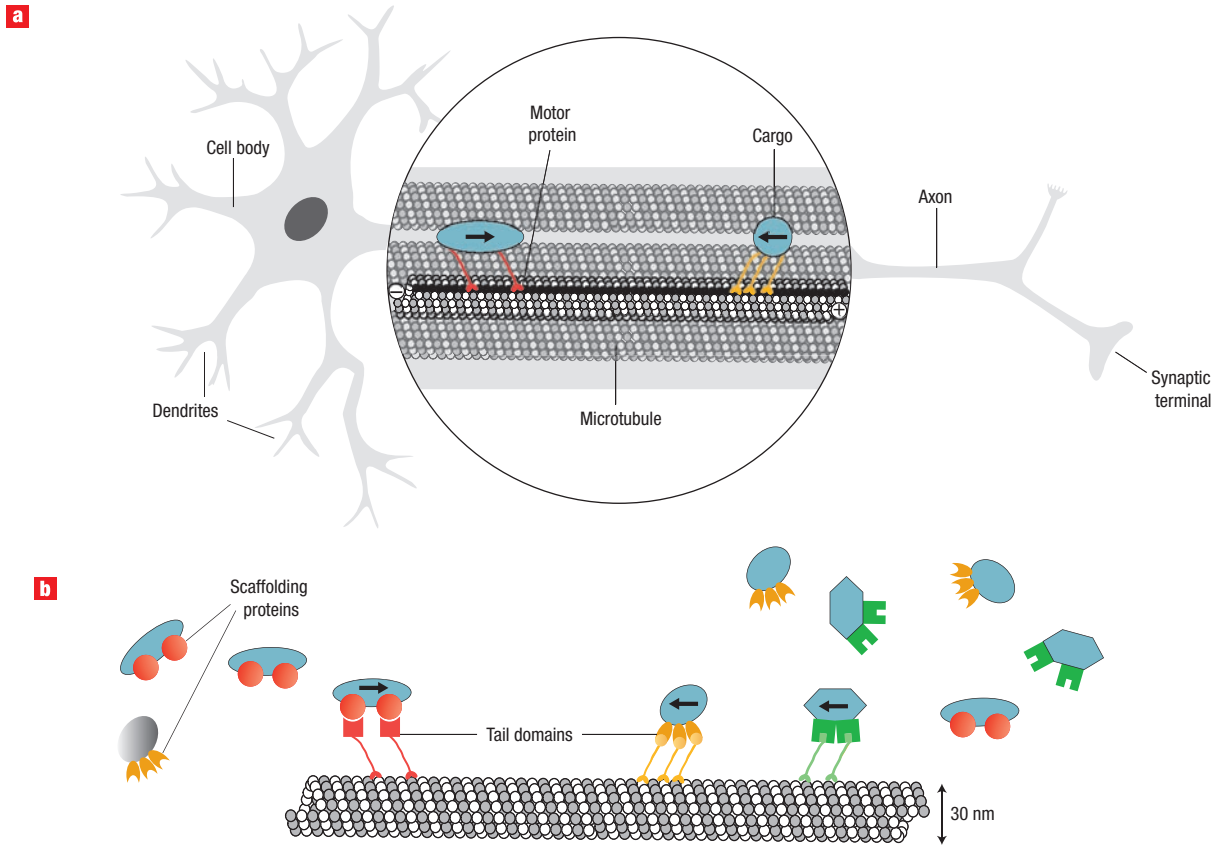


Figure 2 Motor-specific cargo transport in neurons. **a**, The axon of neurons consists of a bundle of highly aligned microtubules along which cargo is trafficked from the cell body to the synapse and vice versa. Most members of the large kinesin family (red) transport cargo towards the periphery, while other motors, including dyneins (yellow), transport cargo in the opposite direction. Motors preferentially move along a protofilament rather then side-stepping (one randomly selected protofilament is shown in dark grey). Protofilaments are assembled from the dimeric protein tubulin (white and grey spheres) which gives microtubules their structural polarity. The protofilaments then form the hollow microtubule rod. When encountering each other on the same protofilament, the much more tightly bound kinesin has the 'right of way', perhaps even forcing the dynein to step sidewise to a neighbouring protofilament[52–55]. **b**, Each member of a motor family selects its own cargo (blue shapes) through specific binding by scaffolding proteins (coloured symbols) or directly by the cargo's tail domains.

Building such copolymers with polymerase nanomotors *ex vivo* would yield much more homogeneous products of the correct sequence and precise length. Natural (for example, nanomotor-enabled) designs could inspire new technologies to synthesize custom biopolymers precisely from a given blueprint.

Ribosome motors have likewise been harnessed *ex vivo* to drive the assembly of new bio-inorganic heterostructures[27] and peptide nanowires[28,29] with gold-modified amino acids inserted into a polypeptide chain. These ribosomes are forced to use inorganically modified t-RNAs to sequentially assemble a hybrid protein containing gold nanoparticles wherever the amino acid cysteine was specified by the messenger RNA template. Such hybrid gold-containing proteins can then attach themselves selectively to materials used in electronics, such as gallium arsenide[28]. This application illustrates how biomotors could be harnessed to synthesize and assemble even non-biological constructs such as nanoelectronic components (see www.cambrios.com).

Assembly nanomotors achieve such high precision in sequential assembly by making use of three key features: (i) geometric shape-fitting selection of their building blocks (for example, nucleotides); (ii) motion along a polymeric template coupled to consumption of an energy source (for example, hydrolysis of ATP molecules); and (iii) intricate proofreading machinery to correct errors as they occur. Furthermore, nanomotor-driven assembly processes allow much more stable, precise and complex nanostructures to be engineered than can be achieved by thermally driven self-assembly techniques alone[30–32].

We should also ask whether some of these principles, which work so well at the nanoscale, could be realized at the micrometre-scale as well. Whitesides and co-workers, for example, have used simple molecular self-assembly strategies, driven by the interplay of hydrophobic and hydrophilic interactions, to assemble microfabricated objects at the mesoscale[33,34]. Perhaps the design principles used by nanomotors to improve precision and correct errors could also be harnessed to engineer future *ex vivo* systems at the nanoscale as well as on other length scales. Learning how to engineer systems that mimic the precision and control of nanomotor-driven assembly processes may ultimately lead to efficient fabrication of complex nanoscopic and mesoscopic structures.

CARGO TRANSPORT

Cells routinely use another set of nanomotors (that is, transport nanomotors) to recognize, sort, shuttle and deliver intracellular cargo along filamentous freeways to well-defined destinations, allowing molecules and organelles to become highly organized (see reviews[35–44]). This is essential for many life processes. Motor

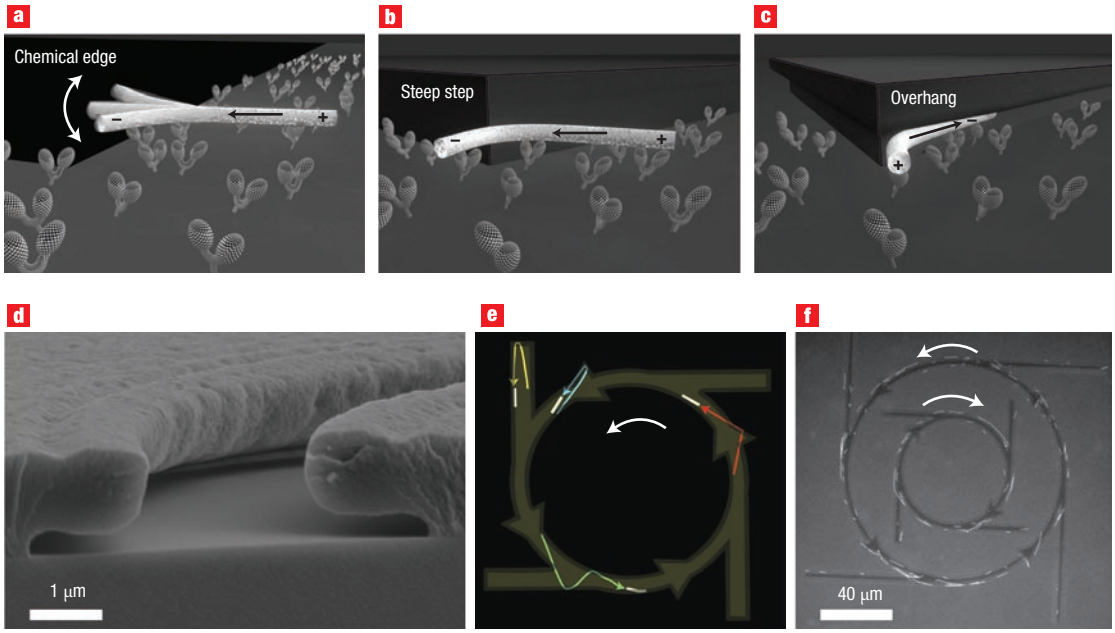


Figure 3 Track designs to guide nanomotor-driven filaments *ex vivo*. A variety of track designs have been used. **a**, A chemical edge (adhesive stripes coated with kinesin surrounded by non-adhesive areas). The filament crosses the chemical edge and ultimately falls off as it does not find kinesins on the non-adhesive areas[61]. **b**, Steep channel walls keep the microtubule on the desired path as they are forced to bend[61,65]. **c**, Overhanging walls have been shown to have the highest guidance efficiency[64]. **d**, Electron micrograph of a microfabricated open channel with overhanging walls[64]. **e**, Breaking the symmetry of micropatterns can promote directional sorting of filament movement[63,65,69,138]. The trajectories of four microtubules are shown: movement into reflector arms causes the tubule to turn around (yellow), an arrow-shaped direction rectifier allows those travelling in the desired direction to continue (red) and forces others to turn around (blue). At intersections tubules preferentially continue straight on (green). **f**, The complex microfabricated circuit analysed in **e** with open channels and overhanging walls demonstrating unidirectional movement of microtubules.

proteins transport cargo along cytoskeletal filaments to precise targets, concentrating molecules in desired locations. In intracellular transport, myosin motors are guided by actin filaments, whereas dynein and kinesin motors move along rod-like microtubules. Figure 2a illustrates how conventional kinesins transport molecular cargo along nerve axons towards the periphery, efficiently transporting material from the cell body to the synaptic region[45]. Dyneins, in contrast, move cargo in the opposite direction, so that there is active communication and recycling between both ends (see reviews[42,46]). In fact, the blockage of such bidirectional cargo transport along nerve axons can give rise to substantial neural disorders[47–50].

The long-range guidance of cargo is made possible by motors pulling their cargo along filamentous rods. Microtubules, for example, are polymerized from the dimeric tubulin into protofilaments that assemble into rigid rods around 30 nm in diameter[36]. These polymeric rods are inherently unstable: they polymerize at one end (plus) while depolymerizing from the other (minus) end, giving rise to a structural polarity. The biological advantage of using transient tracks is that they can be rapidly reconfigured on demand and in response to changing cellular needs or to various external stimuli. Highly efficient unidirectional cargo transport is realized in cells by bundling microtubules into transport highways where all microtubules are oriented in the same direction. Excessively tight bundling of microtubules, however, can greatly impair the efficiency of cargo transport, by blocking the access of motors and cargo to the microtubules in the bundle interior. Instead, microtubule-associated proteins are thought to act as repulsive polymer-brushes, thereby regulating the proximity and interactions between neighbouring microtubules[51].

Traffic control is an issue when using the filaments as tracks on which kinesin and dynein motors move in opposite directions. Although different cargoes can be selectively recognized by different members of the motor protein families and shuttled to different destinations, what happens if motors moving in opposite directions encounter each other on the same protofilament (Fig. 2b)? If two of these motors happen to run into each other, kinesin seems to have the 'right of way'. As kinesin binds the microtubule much more strongly, it is thought to force dynein to step sideways to a neighbouring protofilament[52]. Dynein shows greater lateral movement between protofilaments than kinesin[52–54] as there is a strong diffusional component to its steps[55]. When a microtubule becomes overcrowded with only kinesins, the runs of individual kinesin motors are minimally affected. But when a microtubule becomes overloaded with a mutant kinesin that is unable to step efficiently, the average speed of wild-type kinesin is reduced, whereas its processivity is hardly changed. This suggests that kinesin remains tightly bound to the microtubule when encountering an obstacle and waits until the obstacle unbinds and frees the binding site for kinesin's next step[56].

Engineering principle no. 2: Various track designs enable motors to pull their cargo along filamentous tracks, whereas others allow motors bound to micro- or nanofabricated tracks to propel the filaments which can then serve as carriers.

It is not a trivial task to engineer transport highways *ex vivo*, particularly in versatile geometries with intersections and complex shapes. Individual filaments typically allow only one-dimensional transport, as the motor-linked cargo drops off once the end of the filament is reached. Furthermore, conventional kinesin makes only a few hundred 8-nm-sized steps before dissociating from the microtubule[57,58], further limiting the use of such a system for *ex vivo* applications.

Instead of having the motors transport their cargo along filaments, motors have been immobilized on surfaces in an inverted geometry

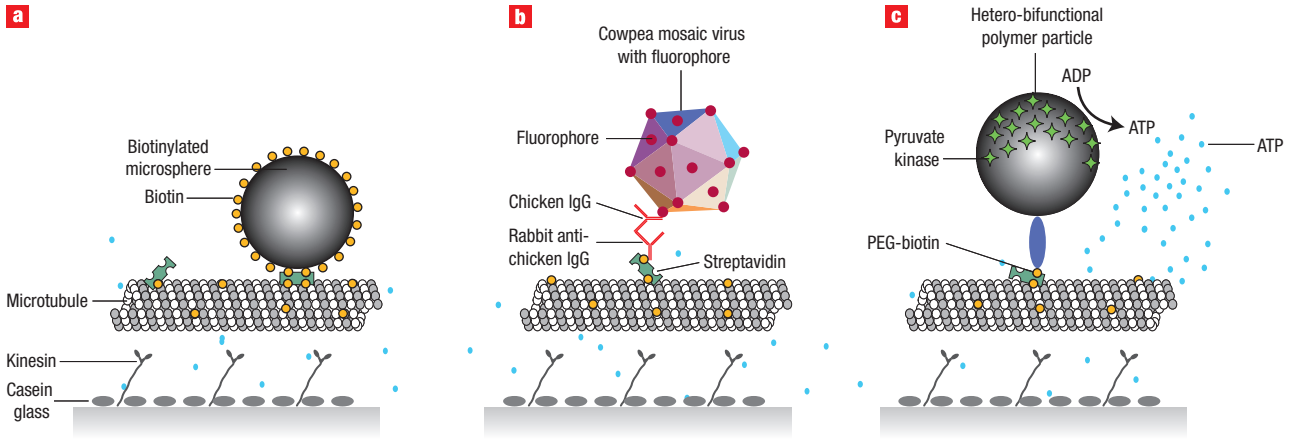


Figure 4 Selecting specific cargo by molecular recognition**.** A versatile toolbox exists by which synthetic and biological cargo can be coupled to microtubules. **a**, Biotinylated objects are coupled via avidin or streptavidin to biotinylated microtubules. **b**, Biological molecules, viruses[79,81] or cells can be coupled by antibody recognition. **c**, Backpacks of chemically or biologically active reagents can be shuttled around, including bioprobes[80] or tiny ATP factories[93] as shown here.

that enables the filaments to be collectively propelled forward[45]. The head domains of the kinesin and myosin motors can rotate and swivel with respect to their feet domains, which are typically bound in random orientations to the surface. These motor heads detect the structural anisotropy of the microtubules and coherently work together to propel a filament forward[59,60].

Various examples of such inverted designs for motor tracks have been engineered to guide filaments efficiently. Some of these are illustrated in Fig. 3. Inverted motility assays can be created, for example, by laying down tracks of motor proteins in microscopic stripes of chemical adhesive on an otherwise flat, protein-repellent surface, surrounded by non-adhesive surface areas. Such chemical patterns (Fig. 3a) have been explored to guide actin filaments or microtubules. The loss rate of guiding filaments increases exponentially with the angle at which they approach an adhesive/non-adhesive contact line[61]. The passage of the contact line by filaments at non-grazing angles, followed by their drop off, can be prevented by using much narrower lanes whose size is of the order of the diameter of the moving object. Such nanoscale kinesin tracks provide good guidance and have been fabricated by nanotemplating[62].

Alternatively, considerably improved guidance has been accomplished by topographic surface features (Fig. 3b). Microtubules hitting a wall are forced to bend along this obstacle and will continue to move along the wall[63–66]. The rigidity of the polymeric filaments used as shuttles thus greatly affects how tracks should be designed for optimal guidance. Whereas microtubules with a persistence length of a few millimetres can be effectively guided in channels a few micrometres wide as they are too stiff to turn around[61], the much more flexible actin filaments require channel widths in the submicrometre range[67,68]. Finally, the best long-distance guidance of microtubules has been obtained so far with overhanging walls[64,69](Fig. 3c). The concept of topographic guidance in fact works so well that swarms of kinesin-driven microtubules have been used as independently moving probes to image unknown surface topographies. After averaging all their trajectories in the focal plane for an extended time period, the image greyscale is determined by the probability of a surface pixel being visited by a microtubule in a given time frame[70].

But how can tracks be engineered to produce *uni*directional cargo transport? All the motor-propelled filaments must move in the same direction to achieve effective long-distance transport. When polar filaments land from solution onto a motor-covered surface, however, their orientations and initial directions of movement are often randomly distributed. Initially, various physical means, such as flow fields[71], have been introduced to promote their alignment. Strong flows eventually either force gliding microtubules to move along with the flow, or force microtubules, if either their plus or minus end is immobilized on a surface[72], to rotate around the anchoring point and along with the flow. The most universal way to control the local direction in which the filamentous shuttles are guided is to make use of asymmetric channel features. Figure 3d–f illustrates how filaments can be actively sorted according to their direction of motion by breaking the symmetry of the engineered tracks. This 'local directional sorting' has been demonstrated on surfaces patterned with open channel geometries, where asymmetric intersections are followed by dead-ended channels (that is, reflector arms), or where channels are broadened into arrow heads. Both of these topographical features not only selectively pass filaments moving in the desired direction, but can also force filaments moving in the opposite direction to turn around[65,69,73,74]. Once directional sorting has been accomplished, electric fields have been used to steer the movement of individual microtubules as they pass through engineered intersections[75,76].

In addition to using isolated nanomotors, hybrid biodevices and systems that harness self-propelling microbes could be used to drive transport processes along engineered tracks. Flagellated bacteria, for example, have been used to generate both translational and rotational motion of microscopic objects[77]. These bacteria can be attached head-on to solid surfaces, either via polystyrene beads or polydimethylsiloxane, thereby enabling the cell bodies to form a densely packed monolayer, while their flagella continue to rotate freely. In fact, a microrotary motor, fuelled by glucose and comprising a 20-μm-diameter silicon dioxide rotor, can be driven along a silicon track by the gliding bacterium *Mycoplasma*[78]. Depending on the specific application and the length scale on which transport needs to be achieved, integrating bacteria into such biohybrid devices (that work under physiological conditions) might ultimately prove more robust than relying solely on individual nanomotors.

CARGO SELECTION

To maintain intracellular contents in an inhomogeneous distribution far from equilibrium, the intracellular transport system must deliver molecular cargo and organelles on demand to precise destinations. This tight spatiotemporal control of molecular deliveries is critical for adequate cell function and survival. Molecular cargo or organelles

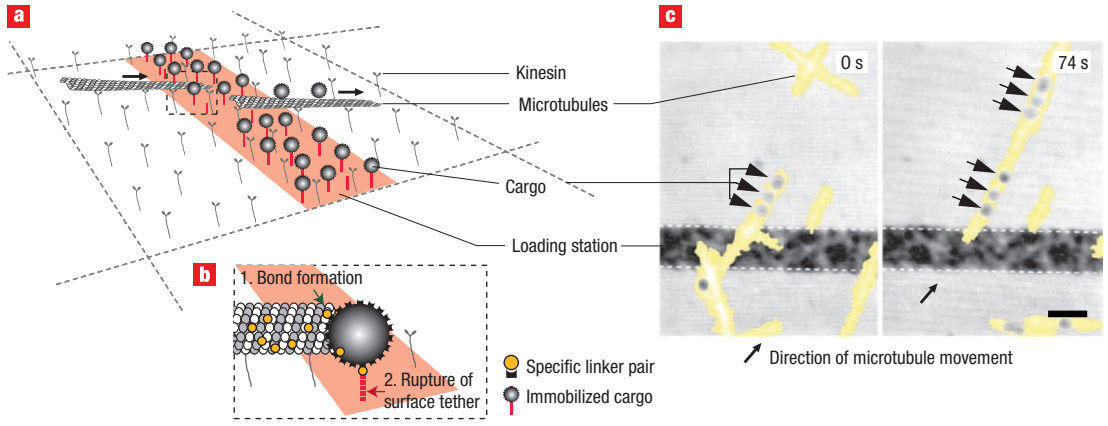


Figure 5 Cargo loading stations[93]. **a**, Stripes of immobilized cargo are fabricated by binding thiolated oligonucleotides to micropatterned lines of gold. Hybridization with complementary strands exposing antibodies at their terminal ends allows them to immobilize a versatile range of cargos that carry antibodies on their surfaces. **b**, The challenge is to tune the bond strength and valency to prevent thermal activation during cargo storage on the loading station. On collision with the shuttle (microtubule), the cargo must rapidly break off the bond it has formed with the station[88]. Fortunately, however, tensile mechanical force acting on a non-covalent bond shortens its lifetime. **c**, **d**, These concepts are used in the design of the loading stations shown here, where a microtubule moves through a stripe of immobilized gold cargo and picks up a few beads.

are typically barcoded so that they can be recognized by their specific motor protein (Fig. 4). Within cells, motors recognize cargo either from the cargo's tail domains directly, or via scaffolding proteins that link cargo to their tail domain[43].

Engineering principle no. 3: Engineered molecular recognition sites enable cargo to be selectively bonded to moving shuttles.

Although most cargo shuttled around by motors can be barcoded using the existing repertoire of biological scaffolding proteins, synthetic approaches are needed for all those *ex vivo* applications where the cargo has to be specifically linked to moving filaments. The loading and transport of biomedically relevant or engineered cargo has already been demonstrated (Fig. 4)[79–83]. Typical approaches are to tag the cargo with antibodies or to biotinylate microtubules and coat the cargo with avidin or streptavidin (Fig. 4) (for reviews, see refs 74,79), as done for polymeric and magnetic beads[84,85] (Fig. 4a), gold nanoparticles[86–88], DNA[87,89,90] and viruses[79,81] (Fig. 4b), and finally mobile bioprobes and sensors[80,81,91](Fig.4c). However, if too much cargo is loaded onto the moving filaments and access of the propelling motors is even partially blocked, the transport velocity can be significantly impaired[92]. Finally, the binding of cargo to a moving shuttle can be used to regulate its performance. In fact, microtubules have recently been furnished with a backpack that self-supplies the energy source ATP. Cargo particles bearing pyruvate kinase have been tethered to the microtubules to provide a local ATP source[93] (Fig. 4c). The coupling of multiple motors to cargo or other scaffold materials can affect the motor performance. If single-headed instead of double-headed kinesins are used, cooperative interactions between the monomeric motors attached to protein scaffolds increase hydrolysis activity and microtubule gliding velocity[59].

At the next level of complexity, successful cargo tagging, sorting and delivery will depend on the engineering of integrated networks of cargo loading, cargo transport and cargo delivery zones. Although the construction of integrated transport circuits is still in its infancy, microfabricated loading stations have been built[88] (Fig. 5). The challenge here is to immobilize cargo on loading stations such that it is not easily detached by thermal motion, yet to allow for rapid cargo transfer to passing microtubules. By properly tuning bond strength and multivalency, and most importantly by taking advantage of the fact that mechanical strain weakens bonds, cargo can be efficiently stored on micropatches and transferred after colliding with a microtubule[88]. Considerable fine-tuning of bond strength can be accomplished by using DNA oligomers hybridized such that the bonds are either broken by force all at once (a strong bond) or in sequence (a weak bond)[94].

As discussed above, filaments are most commonly used to shuttle molecular cargo in most emerging devices that harness linear motors for active transport. Alternatively, if the filamentous tracks could be engineered in versatile geometries, the motors themselves could be used to drag cargo coupled to the molecular recognition sites of their tail domains as in the native systems. We could thus make use of the full biological toolbox of already known or engineered scaffolding proteins that link specific motors to their respective cargoes[40,43]. So far, assemblies of microtubules organized into complex, three-dimensional patterns such as asters, vortices and networks of interconnected poles[95,96] have been successfully created in solution, and mesoscopic needles and rotating spools of microtubule bundles held together by non-covalent interactions have been engineered on surfaces[31]. All of these mesoscopic structures are uniquely related to active motor-driven motion and would not have formed purely by self-assembly without access to an energy source.

To increase the complexity of microtubule track networks, densely packed arrays of microtubules have been grown in confined spaces, consisting of open microfabricated channels with user-defined geometrical patterns[97]. The key to achieving directed transport, however, is for all microtubules within each bundle or array to be oriented in the same direction. This has been accomplished by making use of directed motility in combination with sequential assembly procedures (Fig. 6). First, microtubule seedlings have been oriented in open microfabricated and kinesin-coated channels that contain reflector arms. Once oriented by self-propelled motion, the seedlings were polymerized into mature microtubules that were confined to grow in the open channels until the channels were filled with dense networks of microtubules all oriented in the same direction[97]. Single kinesins take only a few hundred steps before they fall off, but the walking distance can be greatly increased if the cargo is pulled by more than one motor[98]. Such approaches to fabricating networks of microtubule bundles could be further expanded to engineer future devices that use either the full toolbox of native scaffolding proteins or new scaffolding proteins that target both biological and synthetic cargo.

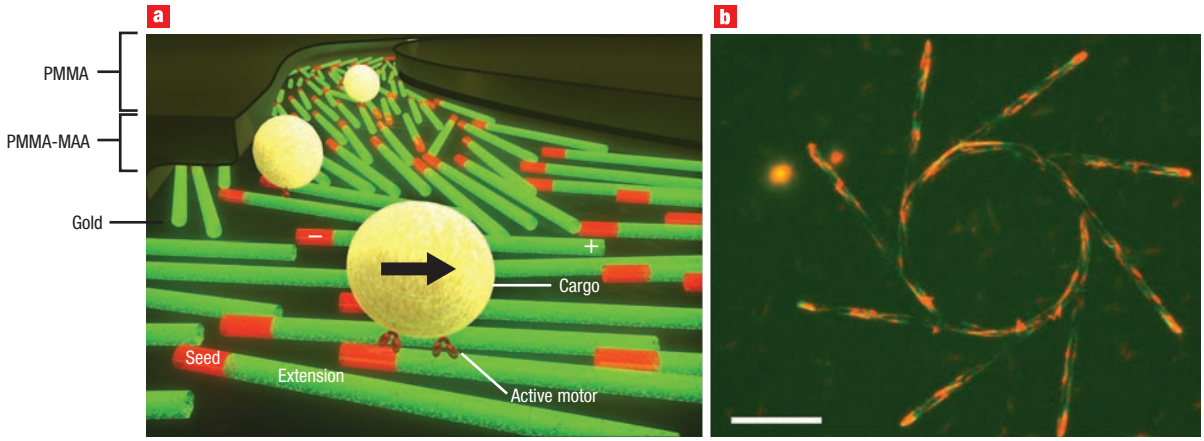


Figure 6 Filament tracks made from engineered bundles of microtubules[97]. Active transport is used to produce bundles of microtubules and confine them to user-defined geometries. **a**, Sequential assembly procedure: first, microtubule seedlings labelled in red are allowed to orient themselves in open kinesin-coated microfabricated channels that contained reflector arms. Second, and after mild fixation, the oriented seedlings are polymerized into mature microtubules through the addition of tubulin into the solution (labelled green) which preferentially binds to the plus-end (polymerizing end) of the microtubules. **b**, Fluorescence image of microtubules that have been grown in the confined space provided by the open channels until the channels were filled with dense networks of microtubules all oriented in the same direction[97]. Scale bar, 40 μm.

Nanoengineers would not be the first to harness biological motors to transport their cargo. Various pathogens are known to hijack microtubule or actin-based transport systems within host cells (reviewed in ref. 99). *Listeria monocytogenes*, for example, propels itself through the host cell cytoplasm by means of a fast-polymerizing actin filament tail[100]. Likewise, the vaccinia virus, a close relative of smallpox, uses actin polymerization to enhance its cell-to-cell spreading[101], and the alpha herpes virus hijacks kinesins to achieve long-distance transport along the microtubules of neuronal axons[102]. Signalling molecules and pathogens that cannot alter cell function and behaviour by simply passing the outer cell membrane can thus hijack the cytoskeletal highways to get transported from the cell periphery to the nucleus.

Engineering principle no. 4: By taking advantage of the existing cytoskeleton, tailored drugs and gene carriers can be actively transported to the cell nucleus.

Indeed, many viruses[37,103,104] as well as non-viral therapeutic gene carriers, such as polyethylenimine/DNA or other polymer-based gene transfer systems (that is, polyplexes)[105,106] take advantage of nanomotor-driven transport along microtubule filaments to accelerate their way through the cytoplasm towards the nucleus. Nanomotor-driven transport to the nucleus leads to a much more efficient nuclear localization than could ever be achieved by slow random diffusion through the viscous cytoplasm. Active gene carrier transport can lead to more efficient perinuclear accumulation within minutes[37,105,106]. In contrast, non-viral gene carriers that depend solely on random diffusion through the cytoplasm move much more slowly and thus have considerably reduced transfection efficiencies. Understanding how to 'hijack' molecular and cellular transport systems, instead of letting a molecule become a target for endosomal degradation[37,91], will ultimately allow the design of more efficient drug and gene carrier systems.

QUALITY CONTROL

Nanomanufacturing processes, much like macroscopic assembly lines, urgently need procedures that offer precise control over the quality of the product, including the ability to recognize and repair defects. Living systems use numerous quality control procedures to detect and repair defects occurring during the synthesis and assembly of biological nanostructures. As yet, this has not been possible in synthetic nanosystems. Many cellular mechanisms for damage surveillance and error correction rely on nanomotors. Such damage control can occur at two different levels as follows.

Engineering principle no. 5: Certain motor proteins recognize assembly mistakes and repair them at the molecular level.

DNA replication represents one of the most complex sequential assembly processes in a cell. Here the genetic information stored in the four-base code must be copied with ultra-high precision. Errors generated during replication can have disastrous biological consequences. Figure 7 illustrates the built-in mechanism used by the polymerase (DNAp) motor to repair mistakes made during the process of DNA replication[107]. When the DNAp motor misincorporates a base while replicating the template DNA strand, it slows down and switches gears from the polymerase to the exonuclease cycle. Once in exonuclease mode, it will excise the mismatched base pair and then rapidly switch back to the polymerase cycle to resume forward replication. Similar error correction mechanisms, known as 'kinetic proofreading', are conjectured to occur in RNA polymerases and ribosomal machineries[1,13,108–113].

Engineering principle no. 6: Integrated systems of motors and signalling molecules are needed to recognize and repair damage at the supramolecular level.

Nerve cells have evolved a highly regulated axonal transport system that contains an integrated damage surveillance system[114]. The traffic regulation of motors moving in opposite directions on a microtubule typically occurs in special 'turnaround' zones at the base and tip of an axon[43], but a zone for switching the organelle's direction can also be created when axonal transport is blocked at the site of nerve injury[46] (see Fig. 2). When irreparable, such blockages are often signatures of neurodegenerative diseases. For example, amyloid precursor protein[47] or tau[115] can give rise to the accumulation of protein aggregates that inhibit anterograde axonal transport, a mechanism potentially implicated in Alzheimer's disease.

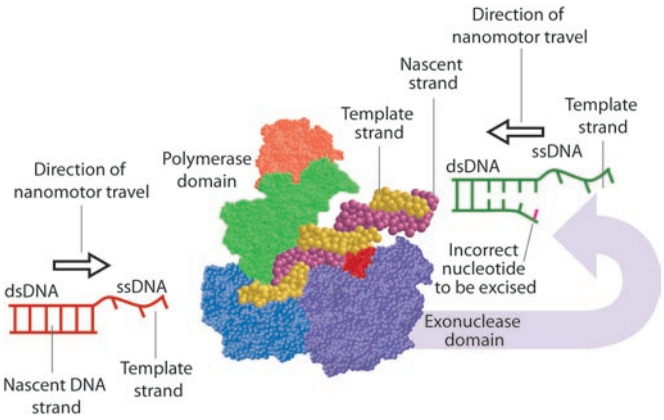


Figure 7 Quality control procedures for damage recognition and molecular repair. The DNA polymerase motor (DNAp) contains two active sites. It switches from polymerase (copying) to exonuclease (error correction) activity when it encounters a mismatched base. Mismatched bases are detected as they have weaker bonding interactions—the 'melting' temperature is lower—and this increases the chance of switching from the polymerase to the exonuclease active site[107]. In the exonuclease mode, the motor excises the incorrect base from the nascent DNA strand.

At present, there are no synthetic materials that can, in a self-regulated manner, recognize and repair defects at either the molecular or supramolecular level. Molecular recognition and repair is typically attributed to a tightly fitted stereochemical complementarity between binding partners. Nanoscale tools applied to the study of molecular recognition and repair are also elucidating the functional roles of the different structural conformations (and hence three-dimensional shapes) of the motors. For instance, the DNAp motor is in one particular conformation when it binds DNA in its copying (that is, polymerization) mode and in an entirely different conformation (that is, the exonuclease mode) when it binds DNA to proofread or excise a mistaken base from the replicated DNA strand[107]. In contrast, damage control at the supramolecular level (for example, during axonal transport) is achieved by the trafficking of signalling molecules. Deciphering the underlying engineering design principles of damage surveillance and error correction mechanisms in biological systems will inevitably allow better quality-control procedures to be integrated into nanoengineered systems.

EXTERNAL CONTROL

Engineering principle no. 7: As with macroscopic engines, external controls can regulate the performance of nanomotors on demand.

Learning how to control and manipulate the performance of nanomotors externally is another critical hurdle in harnessing nanomotors for *ex vivo* applications. By finding or engineering appropriate external knobs in the motor or its environment, its nanoscale movement can be tightly regulated, switched on and off, or otherwise manipulated on demand.

To achieve external control over the nanoscale movement of biological motors, it is important to identify the correct external parameters that can be used to control their dynamics. These external modulators of motor function ('handles') can be either naturally occurring or somehow artificially engineered into the motor to make it susceptible to a particular external control knob or regulator. Because the motion of nanomotors is typically driven by a series of conformational changes in the protein, mechanical load or strain on the motor molecule can also affect the dynamics of the motor. Nanomotors apply mechanical strain to their filaments or substrates as they go through various internal conformational changes. This mechanical strain is intimately related to their dynamics along the substrate and hence their functional performance. Certain interstate transition rates can depend, for example[107], on the amount of intramolecular strain in the motor protein. Applying a mechanical load to a motor perturbs key mechanical transitions in the motor's kinetic pathway, and can thereby affect rates of nucleotide binding, ATP hydrolysis and product release. Single-molecule techniques are beginning to elucidate how mechanical strain on a motor protein might be used to regulate its biological functions (for example, nanoscale assembly or transport)[13,55,107,116–120].

The single-molecule dynamics of the DNA polymerase (DNAp) motor, as it converts single-stranded (ss) DNA to double-stranded (ds) DNA, has been probed, for example, through the differential elasticity of ssDNA and dsDNA (see Fig. 8). The T7 DNA polymerase motor replicates DNA at rates of more than 100 bases per second and this rate steadily decreases with mechanical tension greater than about 5 pN on the DNA template[9]. The motor can work against a maximum of about 34 pN of template tension[9]. The replication rates for the Klenow and Sequenase DNA polymerases also decrease when the ssDNA template tension exceeds 4 pN, and completely ceases at tensions greater than 20 pN (ref. 121). Likewise, single-molecule techniques have allowed direct observation of the RNA polymerase (RNAp) motor moving one base at a time[122], and occasionally pausing and even backtracking[123]. Although RNAp motors are typically five- to tenfold slower than DNAp motors, the effects of DNA template tension on their dynamics are still being investigated[6]. Similarly, ribosome motors, which translate messenger RNA (mRNA) into amino acids at roughly 10 codons per second, have been found to generate about 26.5 ± 1 pN of force[124]. The underlying design principles by which these nanomotors operate are being further elucidated by theoretical models[107,116,125–128] that describe nanomachines at a level commensurate with single-molecule data. Furthermore, these molecular assembly machines can be actively directed, driven and controlled by environmental signals[107].

Consequently, an external load or force applied to the substrate or to the motor itself can be used to slow down a motor's action or stall its movement. The stalling forces of kinesin and dynein are 6 and 1 pN, respectively[58,129]. For example, the binding of two kinesin domains to a microtubule track creates an internal strain in the motor that prevents ATP from binding to the leading motor head. In this way, the two motor domains remain out-of-phase for many mechanochemical cycles and thereby provide an efficient, adaptable mechanism for achieving highly processive movement[130]. Beyond stalling the movement of motors by a mechanical load, other types of perturbations can also influence the dynamics of molecular motors, including the stretching of substrate molecules like DNA[13]. Although this external control over nanomotors has been demonstrated in a few different contexts *ex vivo*, a rich detailed mechanistic understanding of how such external control knobs can modulate the dynamics of the molecular motor is emerging from recent work on the DNA polymerase motor[9,107,116,121,127,128,131].

Remote-controlling the local ATP concentration by the photo-activated release of caged ATP can allow a nanomotor-driven transport system to be accelerated or stopped on demand[84]. External control knobs or regulators can also be engineered into the motors. For instance, point mutations can be introduced into the gene encoding the motor protein, such that it is engineered to respond to light, temperature, pH or other stimuli[43,85]. Engineering light-sensitive switches into nanomotors enables the rate of ATPase[43,132] to be regulated, thereby providing an alternate handle for tuning the motor's speed, even while the ATP concentration is kept constant and high. When additional ATP-consuming enzymes are present in solution, the rate of ATP depletion regulates the distance the shuttles move after being activated by a light pulse and before again coming to a halt[84].

Future applications could require that instead of all the shuttles being moved at the same time, only those in precisely defined locations

be activated, on demand. Some of the highly conserved residues within motors help to determine the motor's ATPase rate[43]. Introducing chemical switches near those locations might provide a handle for chemical manipulation of the motor's speed. In fact, this has already been realized for a rotary motor[132] as well as for a linear kinesin motor, where the insertion of a Ca^{2+}-dependent chemical switch makes the ATPase activity steeply dependent on Ca^{2+} concentrations[133]. In addition to caged ATP, caged peptides that block binding sites could be used to regulate the motility of such systems. Caged peptides derived from the kinesin C-terminus domain have already been used to achieve photo control of kinesin-microtubule motility[134]. Instead of modulating the rate of ATP hydrolysis, the access of microtubules to the motor's head domain can also be blocked in an environmentally controlled manner. In fact, temperature has already been shown to regulate the number of kinesins that are accessible while embedded in a surface-bound film of thermoresponsive polymers[135].

The nanomotor-driven assembly of DNA by the DNA polymerase motor provides an excellent example of how precision control over the nanomotor can be achieved by various external knobs in the motor's environment[107,116,127,128]. The DNAp motor moves along the DNA template by cycling through a given sequence of geometric shape changes. The sequence of shapes or internal states of the nanomachine can be denoted by nodes on a simple network[107,116,127,128]. As illustrated in Fig. 8, this approach elucidates how mechanical tension on a DNA molecule can precisely control (or 'tune') the nanoscale dynamics of the polymerase motor along the DNA track by coupling into key conformational changes of the motor[107].

Macroscopic knobs to precision-control the motor's movement along DNA tracks can be identified by probing how the motor's dynamics vary with each external control knob (varied one at a time). Efforts are currently under way to control even more precisely the movement of these nanomotors along DNA tracks by tightly controlling the parameters in the motor's environment (see www.nanobiosym.com). Concepts of fine-tuning and robustness could also be extended to describe the sensitivity of other nanomotors (modelled as simple biochemical networks) to various external control parameters[107]. Furthermore, such a network approach[107] provides experimentally testable predictions that could aid the design of future molecular-scale manufacturing methods that integrate nanomotor-driven assembly schemes. External control of these nanomotors will be critical in harnessing them for nanoscale manufacturing applications.

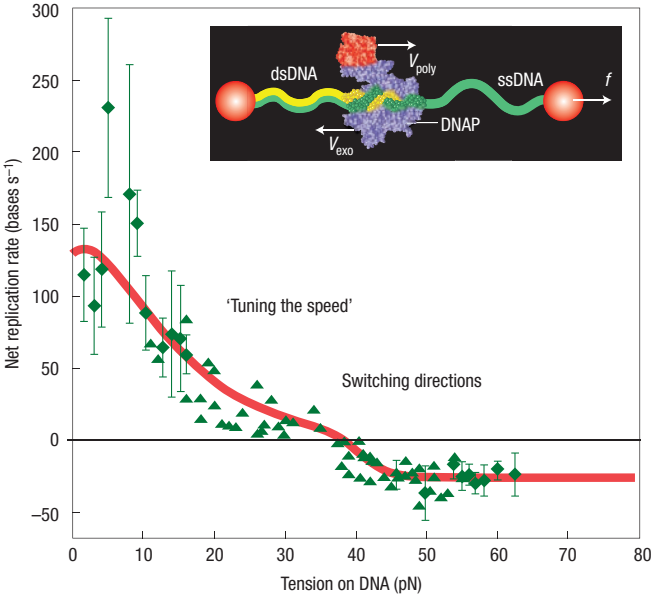


Figure 8 Precision control of nanomotors with external control 'knobs'. The net replication rate of a DNAp motor can be controlled by the mechanical tension on the DNA template strand. Single-molecule data for the motor's force-dependent velocity (two sets of data—diamonds and triangles—are shown, relating to constant force and constant extension measurements) can be described by a network model (red curve) as shown here. The change in net replication rate shows how external controls can change the dynamics of the nanomotor. This model illustrates how environmental control knobs can tune the dynamics of the nanomotor by altering the rate constants associated with its various internal transitions[106]. Tensions between 0 and 35 pN control the net replication rate, whereas tensions above 35 pN actually reverse the velocity of the nanomotor. Inset, experimental setup: a single DNA molecule is stretched between two plastic beads as the motor catalyses the conversion of single-stranded to double-stranded DNA. Figure adapted from ref. 106.

CONCLUDING REMARKS

We have reviewed several key engineering design principles that enable nanomotors moving along linear templates to perform a myriad of tasks. Equally complex biomimetic tasks have not yet been mastered *ex vivo*, either by harnessing biological motors or via synthetic analogues. Engineering insights into how such tasks are carried out by the biological nanosystems will inspire new technologies that harness nanomotor-driven processes to build new systems for nanoscale transport and assembly.

Sequential assembly and nanoscale transport, combined with features currently attributed only to biological materials, such as self-repair and healing, might one day become an integral part of future materials and bio-hybrid devices. In the near term, molecular biology techniques could be used to synthesize and assemble nanoelectronic components with more control (www.cambrios.com; see also ref. 29). Numerous proof-of-concept experiments using nanomotors integrated into synthetic microdevices have already been demonstrated (see reviews[74,136]). Among many others, these applications include stretching surface-bound molecules by moving microtubules[87,90]; probing the lifetime of a single receptor–ligand interaction via a cantilevered microtubule that acts as a piconewton force sensor[85]; topographic surface imaging by self-propelled probes[70]; and cargo pick-up from loading stations[88] as illustrated in Fig. 5.

Although much progress is being made in the synthesis of artificial motors (see review[137]), it has been difficult, in practice, to synthesize artificial motors that come even close in performance to their natural counterparts (see review[39]). Harnessing biological motors to perform nanoscale manufacturing tasks might thus be the best near-term strategy. Although many individual nanoparts can be easily manufactured, the high-throughput assembly of these nanocomponents into complex structures is still non-trivial. At present, no *ex vivo* technology exists that can actively guide such nanoscale assembly processes. Despite advances in deciphering the underlying engineering design principles of nanomotors, many hurdles still impede harnessing them for *ex vivo* transport and sequential assembly in nanosystems. Although the use of biological nanomotors puts intrinsic constraints on the conditions under which they can be assembled and used in biohybrid devices, many of their sophisticated tasks are still poorly mimicked by synthetic analogues. Understanding the details of how these little nanomachines convert chemical energy into controlled movements will nevertheless inspire new approaches to engineer synthetic counterparts that might some day be used under harsher conditions, operate at more extreme temperatures, or simply have longer shelf lives.

Certain stages of the materials production process might one day be replaced by nanomotor-driven sequential self-assembly, allowing much more control at the molecular level. Biological motors are already

being used to drive the efficient fabrication of complex nanoscopic and mesoscopic structures, such as nanowires[31] and supramolecular assemblies. Techniques for precision control of nanomotors that read DNA are also being used to engineer integrated systems for rapid DNA detection and analysis (www.nanobiosym.com). The specificity and control of assembly and transport shown by biological systems offers many opportunities to those interested in assembly of complex nanosystems. Most importantly, the intricate schemes of proofreading and damage repair—features that have not yet been realized in any manmade nanosystems—should provide inspiration for those interested in producing synthetic systems capable of similarly complex tasks.

doi:10.1038/nnano.2008.190

Published online: 27 July 2008.

References

1. Rodnina, M. V. & Wintermeyer, W. Fidelity of aminoacyl-tRNA selection on the ribosome: kinetic and structural mechanisms. *Annu. Rev. Biochem.* **70,** 415–435 (2001).
2. Kunkel, T. A. DNA replication fidelity. *J. Biol. Chem.* **279,** 16895–16898 (2004).
3. Erie, D. A., Hajiseyedjavadi, O., Young, M. C. & von Hippel, P. H. Multiple RNA polymerase conformations and GreA: control of the fidelity of transcription. *Science* **262,** 867 (1993).
4. Liu, D. R., Magliery, T. J., Pastrnak, M. & Schultz, P. G. Engineering a tRNA and aminoacyl-tRNA synthetase for the site-specific incorporation of unnatural amino acids into proteins in vivo. *Proc. Natl Acad. Sci.* **94,** 10092–10097 (1997).
5. Bustamante, C., Smith, S. B., Liphardt, J. & Smith, D. Single-molecule studies of DNA mechanics. *Curr. Opin. Struct. Biol.* **10,** 279–285 (2000).
6. Davenport, R. J., Wuite, G. J. L., Landick, R. & Bustamante, C. Single-molecule study of transcriptional pausing and arrest by *E. coli* RNA polymerase. *Science* **287,** 2497–2500 (2000).
7. Greulich, K. O. Single-Molecule Studies on DNA and RNA. *ChemPhysChem* **6,** 2459–2471 (2005).
8. Wang, M. D. *et al.* Force and velocity measured for single molecules of RNA polymerase. *Science* **282,** 902–907 (1998).
9. Wuite, G. J., Smith, S. B., Young, M., Keller, D. & Bustamante, C. Single-molecule studies of the effect of template tension on T7 DNA polymerase activity. *Nature* **404,** 103–106 (2000).
10. Smith, S. B., Cui, Y. & Bustamante, C. Overstretching B-DNA: The elastic response of individual double-stranded and single-stranded DNA molecules. *Science* **271,** 795 (1996).
11. Smith, S. B., Finzi, L. & Bustamante, C. Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science* **258,** 1122 (1992).
12. Williams, M. C. & Rouzina, I. Force spectroscopy of single DNA and RNA molecules. *Curr. Opin. Struct. Biol* **12,** 330–336 (2002).
13. Bustamante, C., Bryant, Z. & Smith, S. B. Ten years of tension: single-molecule DNA mechanics. *Nature* **421,** 423–427 (2003).
14. Jeney, S., Stelzer, E. H., Grubmuller, H. & Florin, E. L. Mechanical properties of single motor molecules studied by three-dimensional thermal force probing in optical tweezers. *ChemPhysChem* **5,** 1150–1158 (2004).
15. Mehta, A. D. Single-molecule biomechanics with optical methods. *Science* **283,** 1689–1695 (1999).
16. Mogilner, A. & Oster, G. Polymer motors: pushing out the front and pulling up the back. *Curr. Biol.* **13,** 721–733 (2003).
17. Schnitzer, M. J., Visscher, K. & Block, S. M. Force production by single kinesin motors. *Nature Cell Biol.* **2,** 718–723. (2000).
18. Strick, T., Allemand, J. F., Croquette, V. & Bensimon, D. The manipulation of single biomolecules. *Phys. Today* **54,** 46–51 (October, 2001).
19. Ha, T. Single-molecule fluorescence methods for the study of nucleic acids. *Curr. Opin. Struct. Biol.* **11,** 287–292 (2001).
20. Kapanidis, A. N. *et al.* Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism. *Science* **314,** 1144–1147 (2006).
21. Keller, R. A. *et al.* Single-molecule fluorescence analysis in solution. *Appl. Spectrosc.* **50,** 12A-32A (1996).
22. Mullis, K. B. *The Polymerase Chain Reaction*. Nobel Lecture (1993).
23. van Hest, J. C. M. & Tirrell, D. A. Protein-based materials, toward a new level of structural control. *Chem. Commun.* **19,** 1897–1904 (2001).
24. Fodor, S. P. *et al.* Multiplexed biochemical assays with biological chips. *Nature* **364,** 555–556 (1993).
25. Merrifield, R. B. Automated synthesis of peptides. *Science* **150,** 178–185 (1965).
26. Ratner, D. M., Swanson, E. R. & Seeberger, P. H. Automated synthesis of a protected N-linked glycoprotein core pentasaccharide. *Org. Lett.* **5,** 4717–4720 (2003).
27. Ball, P. It all falls into place. *Nature* **413,** 667–668 (2001).
28. Pavel, I. S. *Assembly of gold nanoparticles by ribosomal molecular machines* PhD thesis, Univ. Texas at Austin (2005).
29. Whaley, S. R., English, D. S., Hu, E. L., Barbara, P. F. & Belcher, A. M. Selection of peptides with semiconductor binding specificity for directed nanocrystal assembly. *Nature* **405,** 665–668 (2000).
30. Chen, H. L. & Goel, A. in *DNA Computing. Lecture Notes in Computer Science* Vol. 3384, 62–75 (Springer, Berlin/Heidelberg, 2005).
31. Hess, H. *et al.* Molecular self-assembly of 'nanowires'and 'nanospools' using active transport. *Nano Lett.* **5,** 629–633 (2005).
32. Winfree, E. & Bekbolatov, R. in *DNA Computing. Lecture Notes in Computer Science* Vol. 2943, 126–144 (Springer, Berlin/Heidelberg, 2004).
33. Choi, I. S., Bowden, N. & Whitesides, G. M. Macroscopic, hierarchical, two-dimensional self-assembly. *Angew. Chem. Intl Edn Engl.* **38,** 3078–3081 (1999).
34. Whitesides, G. M. & Boncheva, M. Beyond molecules: self-assembly of mesoscopic and macroscopic components. *Proc. Natl Acad. Sci. USA* **99,** 4769–4774 (2002).
35. Caviston, J. P. & Holzbaur, E. L. Microtubule motors at the intersection of trafficking and transport. *Trends Cell Biol.* **16,** 530–537 (2006).
36. Howard, J. *Mechanics of Motor Proteins and the Cytoskeleton* (Sinauer, Sunderland, Massachusetts, 2001).
37. Lakadamyali, M., Rust, M. & Zhuang, X. Ligands for clathrin-mediated endocytosis are differentially sorted into distinct populations of early endosomes. *Cell* **124,** 997–1009 (2006).
38. Lakadamyali, M., Rust, M. J., Babcock, H. P. & Zhuang, X. Visualizing infection of individual influenza viruses. *Proc. Natl Acad. Sci. USA* **100,** 9280–9285 (2003).
39. Månsson, A. & Linke, H. *Controlled Nanoscale Motion*. Proc. Nobel Symp. 131, Vol. 711. (Springer, Berlin, 2007).
40. Miki, H., Okada, Y. & Hirokawa, N. Analysis of the kinesin superfamily: insights into structure and function. *Trends Cell Biol.* **15,** 467–476 (2005).
41. Rust, M. J., Lakadamyali, M., Zhang, F. & Zhuang, X. Assembly of endocytic machinery around individual influenza viruses during viral entry. *Nature Struct. Mol. Biol.* **11,** 567–573 (2004).
42. Sotelo-Silveira, J. R., Calliari, A., Kun, A., Koenig, E. & Sotelo, J. R. RNA trafficking in axons. *Traffic* **7,** 508–515 (2006).
43. Vale, R. D. The molecular motor toolbox for intracellular transport. *Cell* **112,** 467–480. (2003).
44. Vallee, R. B. & Sheetz, M. P. Targeting of motor proteins. *Science* **271,** 1539–1544 (1996).
45. Vale, R. D., Reese, T. S. & Sheetz, M. P. Identification of a novel force-generating protein, kinesin, involved in microtubule-based motility. *Cell* **42,** 39–50 (1985).
46. Guzik, B. W. & Goldstein, L. S. Microtubule-dependent transport in neurons: steps towards an understanding of regulation, function and dysfunction. *Curr. Opin. Cell Biol.* **16,** 443–450 (2004).
47. Gunawardena, S. & Goldstein, L. S. Disruption of axonal transport and neuronal viability by amyloid precursor protein mutations in Drosophila. *Neuron* **32,** 389–401 (2001).
48. Gunawardena, S. & Goldstein, L. S. Cargo-carrying motor vehicles on the neuronal highway: transport pathways and neurodegenerative disease. *J. Neurobiol.* **58,** 258–271 (2004).
49. Mandelkow E, M. E. Kinesin motors and disease. *Trends Cell Biol.* **12,** 585–591 (2002).
50. Ström, A. L. *et al.* Retrograde axonal transport and motor neuron disease. *J. Neurochem.* Preprint at <http://www.ncbi.nlm.nih.gov/pubmed/18384644> (23 April 2008).
51. Mukhopadhyay, R. & Hoh, J. H. AFM force measurements on microtubule-associated proteins: the projection domain exerts a long-range repulsive force. *FEBS Lett.* **505,** 374–378 (2001).
52. Mizuno, N. *et al.* Dynein and kinesin share an overlapping microtubule-binding site. *EMBO J.* **23,** 2459–2467 (2004).
53. Vale, R. D. & Toyoshima, Y. Y. Rotation and translocation of microtubules in vitro induced by dyneins from Tetrahymena cilia. *Cell* **52,** 459–469 (1988).
54. Wang, Z., Khan, S. & Sheetz, M. P. Single cytoplasmic dynein molecule movements: characterization and comparison with kinesin. *Biophys. J.* **69,** 2011–2023 (1995).
55. Reck-Peterson, S. L. *et al.* Single-molecule analysis of dynein processivity and stepping behavior. *Cell* **126,** 335–348 (2006).
56. Seitz, A. & Surrey, T. Processive movement of single kinesins on crowded microtubules visualized using quantum dots. *EMBO J.* **25,** 267–277 (2006).
57. Coppin, C. M., Finer, J. T., Spudich, J. A. & Vale, R. D. Detection of sub-8-nm movements of kinesin by high-resolution optical-trap microscopy. *Proc. Natl Acad. Sci. USA* **93,** 1913–1917 (1996).
58. Svoboda, K., Schmidt, C. F., Schnapp, B. J. & Block, S. M. Direct observation of kinesin stepping by optical trapping interferometry. *Nature* **365,** 721–727 (1993).
59. Diehl, M. R., Zhang, K., Lee, H. J. & Tirrell, D. A. Engineering cooperativity in biomotor-protein assemblies. *Science* **311,** 1468–1471 (2006).
60. Hunt, A. J. & Howard, J. Kinesin swivels to permit microtubule movement in any direction. *Proc. Natl Acad. Sci. USA* **90,** 11653–11657 (1993).
61. Clemmens, J. *et al.* Principles of microtubule guiding on microfabricated kinesin-coated surfaces: chemical and topographic surface patterns. *Langmuir* **19,** 10967–10974 (2003).
62. Reuther, C., Hajdo, L., Tucker, R., Kasprzak, A. A. & Diez, S. Biotemplated nanopatterning of planar surfaces with molecular motors. *Nano Lett.* **6,** 2177–2183 (2006).
63. Clemmens, J., Hess, H., Howard, J. & Vogel, V. Analysis of microtubule guidance by microfabricated channels coated with kinesin. *Langmuir* **19,** 1738–1744 (2003).
64. Hess, H. *et al.* Molecular shuttles operating undercover: a new photolithographic approach for the fabrication of structured surfaces supporting directed motility. *Nano Lett.* **3,** 1651–1655 (2003).
65. Hiratsuka, Y., Tada, T., Oiwa, K., Kanayama, T. & Uyeda, T. Q. Controlling the direction of kinesin-driven microtubule movements along microlithographic tracks. *Biophys. J.* **81,** 1555–1561. (2001).
66. Moorjani, S. G., Jia, L., Kackson, T. N. & Hancock, W. O. Lithographically patterned channels spatially segregate kinesin motor activity and effectively guide microtubule movements. *Nano Lett.* **3,** 633–637 (2003).
67. Bunk, R. *et al.* Actomyosin motility on nanostructured surfaces. *Biochem. Biophys. Res. Commun.* **301,** 783–788 (2003).
68. Sundberg, M. *et al.* Actin filament guidance on a chip: toward high-throughput assays and lab-on-a-chip applications. *Langmuir* **22,** 7286–7295 (2006).
69. Clemmens, J. *et al.* Motor-protein 'roundabouts': microtubules moving on kinesin-coated tracks through engineered networks. *Lab Chip* **4,** 83–86 (2004).
70. Hess, H., Clemmens, J., Howard, J. & Vogel, V. Surface imaging by self-propelled nanoscale probes. *Nano Lett.* **2,** 113–116 (2002).
71. Stracke, R., Böhm, K. J., Burgold, J., Schacht, H.-J. & Unger, E. Physical and technical parameters determining the functioning of a kinesin-based cell-free motor system. *Nanotechnology* **11,** 52–56 (2000).
72. Brown, T. B. & Hancock, W. O. A polarized microtubule array for kinesin-powered nanoscale assembly and force generation. *Nano Lett.* **28**, 571–576 (2005).
73. Nitta, T., Tanahashi, A., Hirano, M. & Hess, H. Simulating molecular shuttle movements: towards computer-aided design of nanoscale transport systems. *Lab Chip* **6,** 881–885 (2006).
74. Vogel, V. & Hess, H. in *Lecture Notes Proceedings Nobel Symposium* Vol. 711, 367–383 (Springer, Berlin/Heidelberg, 2007).

75. Stracke, R., Bohm, K. J., Wollweber, L., Tuszynski, J. A. & Unger, E. Analysis of the migration behaviour of single microtubules in electric fields. *Biochem. Biophys. Res. Commun.* **293,** 602–609. (2002).
76. van den Heuvel, M. G., de Graaff, M. P. & Dekker, C. Molecular sorting by electrical steering of microtubules in kinesin-coated channels. *Science* **312,** 910–914 (2006).
77. Darnton, N., Turner, L., Breuer, K. & Berg, H. C. Moving fluid with bacterial carpets. *Biophys. J.* **86,** 1863–1870 (2004).
78. Hiratsuka, Y., Miyata, M., Tada, T. & Uyeda, T. Q. A microrotary motor powered by bacteria. *Proc. Natl Acad. Sci. USA* **103,** 13618–13623 (2006).
79. Bachand, G. D., Rivera, S. B., Carroll-Portillo, A., Hess, H. & Bachand, M. Active capture and transport of virus particles using a biomolecular motor-driven, nanoscale antibody sandwich assay. *Small* **2,** 381–385 (2006).
80. Hirabayashi, M. *et al.* Malachite green-conjugated microtubules as mobile bioprobes selective for malachite green aptamers with capturing/releasing ability. *Biotechnol. Bioeng.* **94,** 473–480 (2006).
81. Martin, B. D. *et al.* An engineered virus as a bright fluorescent tag and scaffold for cargo proteins: capture and transport by gliding microtubules. *J. Nanosci. Nanotechnol.* **6,** 2451–2460 (2006).
82. Muthukrishnan, G., Hutchins, B. M., Williams, M. E. & Hancock, W. O. Transport of semiconductor nanocrystals by kinesin molecular motors. *Small* **2,** 626–630 (2006).
83. Taira, S. *et al.* Selective detection and transport of fully matched DNA by DNA-loaded microtubule and kinesin motor protein. *Biotechnol. Bioeng.* **95,** 533–538 (2006).
84. Hess, H., Clemmens, J., Qin, D., Howard, J. & Vogel, V. Light-controlled molecular shuttles made from motor proteins carrying cargo on engineered surfaces. *Nano Lett.* **1,** 235–239 (2001).
85. Hess, H., Howard, J. & Vogel, V. A piconewton forcemeter assembled from microtubules and kinesins. *Nano Lett.* **2,** 1113–1115 (2002).
86. Boal, A. K., Bachand, G. D., Rivera, S. B. & Bunker, B. C. Interactions between cargo-carrying biomolecular shuttles. *Nanotechnology* **17,** 349–354 (2006).
87. Diez, S. *et al.* Stretching and transporting DNA molecules using motor proteins. *Nano Lett.* **3,** 1251–1254 (2003).
88. Brunner, C., Wahnes, C. & Vogel, V. Cargo pick-up from engineered loading stations by kinesin driven molecular shuttles. *Lab on a Chip* **7,** 1263–1271 (2007).
89. Ramachandran, S., Ernst, K. H., Bachand, G. D., Vogel, V. & Hess, H. Selective loading of kinesin-powered molecular shuttles with protein cargo and its application to biosensing. *Small* **2,** 330 (2006).
90. Dinu, C. Z. *et al.* Parallel manipulation of bifunctional DNA molecules on structured surfaces using kinesin-driven microtubules. *Small* **2,** 1090–1098 (2006).
91. Soldati, T. & Schliwa, M. Powering membrane traffic in endocytosis and recycling. *Nature Rev. Mol. Cell Biol.* **7,** 897–908 (2006).
92. Bachand, M., Trent, A. M., Bunker, B. C. & Bachand, G. D. Physical factors affecting kinesin-based transport of synthetic nanoparticle cargo. *J. Nanosci. Nanotechnol.* **5,** 718–722 (2005).
93. Du, Y. Z. *et al.* Motor protein nano-biomachine powered by self-supplying ATP. *Chem. Commun.*, 2080–2082 (2005).
94. Kufer, S. K., Puchner, E. M., Gumpp, H., Liedl, T. & Gaub, H. E. Single-molecule cut-and-paste surface assembly. *Science* **319,** 594–596 (2008).
95. Chakravarty, A., Howard, L. & Compton, D. A. A mechanistic model for the organization of microtubule asters by motor and non-motor proteins in a mammalian mitotic extract. *Mol. Biol. Cell* **15,** 2116–2132 (2004).
96. Surrey, T., Nedelec, F., Leibler, S. & Karsenti, E. Physical properties determining self-organization of motors and microtubules. *Science* **292,** 1167–1171. (2001).
97. Doot, R. K., Hess, H., Vogel, V. Engineered networks of oriented microtubule filaments for directed cargo transport. *Soft Matter* **3,** 349 - 356 (2007).
98. Beeg, J. *et al.* Transport of beads by several kinesin motors. *Biophys. J.* **94,** 532 (2008).
99. Henry, T., Gorvel, J. P. & Meresse, S. Molecular motors hijacking by intracellular pathogens. *Cell Microbiol.* **8,** 23–32 (2006).
100. Soo, F. S. & Theriot, J. A. Large-scale quantitative analysis of sources of variation in the actin polymerization-based movement of *Listeria monocytogenes*. *Biophys. J.* **89,** 703–723 (2005).
101. Rietdorf, J. *et al.* Kinesin-dependent movement on microtubules precedes actin-based motility of vaccinia virus. *Nature Cell Biol.* **3,** 992–1000 (2001).
102. Smith, G. A., Gross, S. P. & Enquist, L. W. Herpes viruses use bidirectional fast-axonal transport to spread in sensory neurons. *Proc. Natl Acad. Sci. USA* **98,** 3466–3470 (2001).
103. Döhner, K., Nagel, C.-H. & Sodeik, B. Viral stop-and-go along microtubules: taking a ride with dynein and kinesins. *Trends Microbiol.* **13,** 320–327 (2005).
104. Sodeik, B. Unchain my heart, baby let me go: the entry and intracellular transport of HIV. *Cell Biol.* **159,** 393–395 (2002).
105. Kulkarni, R. P., Wu, D. D., Davis, M. E. & Fraser, S. E. Quantitating intracellular transport of polyplexes by spatio-temporal image correlation spectroscopy. *Proc. Natl Acad. Sci. USA* **102,** 7523–7528 (2005).
106. Suh, J., Wirtz, D. & Hanes, J. Efficient active transport of gene nanocarriers to the cell nucleus. *Proc. Natl Acad. Sci. USA* **100,** 3878–3882. (2003).
107. Goel, A., Astumian, R. D. & Herschbach, D. Tuning and switching a DNA polymerase motor with mechanical tension. *Proc. Natl Acad. Sci. USA* **100,** 9699–9704 (2003).
108. Donlin, M. J., Patel, S. S. & Johnson, K. A. Kinetic partitioning between the exonuclease and polymerase sites in DNA error correction. *Biochemistry* **30,** 538–546 (1991).
109. Fersht, A. R., Knill-Jones, J. W. & Tsui, W. C. Kinetic basis of spontaneous mutation. Misinsertion frequencies, proofreading specificities and cost of proofreading by DNA polymerases of *Escherichia coli*. *J. Mol. Biol.* **156,** 37–51 (1982).
110. Hopfield, J. J. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Natl Acad. Sci. USA* **71,** 4135–4139 (1974).
111. Hopfield, J. J. The energy relay: a proofreading scheme based on dynamic cooperativity and lacking all characteristic symptoms of kinetic proofreading in DNA replication and protein synthesis. *Proc. Natl Acad. Sci. USA* **77,** 5248 (1980).
112. Rodnina, M. V. & Wintermeyer, W. Ribosome fidelity: tRNA discrimination, proofreading and induced fit. *Trends Biochem. Sci.* **26,** 124–130 (2001).
113. Wang, D. & Hawley, D. K. Identification of a 3'→5'exonuclease activity associated with human RNA polymerase II. *Proc. Natl Acad. Sci. USA* **90,** 843–847 (1993).
114. Cavalli, V., Kujala, P., Klumperman, J. & Goldstein, L. S. Sunday Driver links axonal transport to damage signaling. *J. Cell Biol.* **168,** 775–787 (2005).
115. Mandelkow, E. M., Stamer, K., Vogel, R., Thies, E. & Mandelkow, E. Clogging of axons by tau, inhibition of axonal traffic and starvation of synapses. *Neurobiol. Aging* **24,** 1079–1085 (2003).
116. Goel, A., Ellenberger, T., Frank-Kamenetskii, M. D. & Herschbach, D. Unifying themes in DNA replication: reconciling single molecule kinetic studies with structural data on DNA polymerases. *J. Biomol. Struct. Dyn.* **19,** 571–584 (2002).
117. Guydosh, N. R. & Block, S. M. Backsteps induced by nucleotide analogs suggest the front head of kinesin is gated by strain. *Proc. Natl Acad. Sci. USA* **103,** 8054–8059 (2006).
118. Spudich, J. Molecular motors take tension in stride. *Cell* **126,** 242–244 (2006).
119. Vale, R. D. & Milligan, R. A. The way things move: looking under the hood of molecular motor proteins. *Science* **288,** 88–95 (2000).
120. Veigel, C., Schmitz, S., Wang, F. & Sellers, J. R. Load-dependent kinetics of myosin-V can explain its high processivity. *Nature Cell Biol.* **7,** 861–869 (2005).
121. Maier, B., Bensimon, D. & Croquette, V. Replication by a single DNA polymerase of a stretched single-stranded DNA. *Proc. Natl Acad. Sci. USA* **97,** 12002–12007 (2000).
122. Abbondanzieri, E. A., Greenleaf, W. J., Shaevitz, J. W., Landick, R. & Block, S. M. Direct observation of base-pair stepping by RNA polymerase. *Nature* **438,** 460–465 (2005).
123. Shaevitz, J. W., Abbondanzieri, E. A., Landick, R. & Block, S. Backtracking by single RNA polymerase molecules observed at near-base pair resolution. *Nature* **426,** 684–687 (2003).
124. Sinha, D. K., Bhalla, U. S. & Shivashankar, G. V. Kinetic measurement of ribosome motor stalling force. *Appl. Phys. Lett.* **85,** 4789–4791 (2004).
125. Astumian, R. D. Thermodynamics and kinetics of a Brownian motor. *Science* **276,** 917–922 (1997).
126. Bustamante, C., Keller, D. & Oster, G. The physics of molecular motors. *Acc. Chem. Res.* **34,** 412–420 (2001).
127. Goel, A., Frank-Kamenetskii, M. D., Ellenberger, T. & Herschbach, D. Tuning DNA 'strings': modulating the rate of DNA replication with mechanical tension. *Proc. Natl Acad. Sci. USA* **98,** 8485–8489 (2001).
128. Goel, A. & Herschbach, D. R. Controlling the speed and direction of molecular motors that replicate DNA. *Proc. SPIE* **5110,** 63–68 (2003).
129. Mallik, R., Carter, B. C., Lex, S. A., King, S. J. & Gross, S. P. Cytoplasmic dynein functions as a gear in response to load. *Nature* **427,** 649–652 (2004).
130. Rosenfeld, S. S., Fordyce, P. M., Jefferson, G. M., King, P. H. & Block, S. M. Stepping and stretching. How kinesin uses internal strain to walk processively. *J. Biol. Chem.* **278,** 18550–18556 (2003).
131. Andricioaei, I., Goel, A., Herschbach, D. & Karplus, M. Dependence of DNA polymerase replication rate on external forces: a model based on molecular dynamics simulations. *Biophys. J.* **87,** 1478–1497 (2004).
132. Liu, H. *et al.* Control of a biomolecular motor-powered nanodevice with an engineered chemical switch. *Nature Mater.* **1,** 173–177 (2002).
133. Konishi, K., Uyeda, T. Q. & Kubo, T. Genetic engineering of a Ca(2+) dependent chemical switch into the linear biomotor kinesin. *FEBS Lett.* **580,** 3589–3594 (2006).
134. Nomura, A., Uyeda, T. Q., Yumoto, N. & Tatsu, Y. Photo-control of kinesin-microtubule motility using caged peptides derived from the kinesin C-terminus domain. *Chem. Comm.* **1,** 3588–3590 (2006).
135. Ionov, L., Stamm, M. & Diez, S. Reversible switching of microtubule motility using thermoresponsive polymer surfaces. *Nano Lett.* **6,** 1982–1987 (2006).
136. van den Heuvel, M. G. L. & Dekker, C. Motor proteins at work for nanotechnology. *Science* **317,** 333 - 336 (2007).
137. Browne, W. R. & Feringa, B. L. Making molecular machines work. *Nature Nanotech.* **1,** 25–35 (2006).
138. Hess, H. *et al.* Ratchet patterns sort molecular shuttles. *Appl. Phys. A* **75,** 309–313 (2002).

Acknowledgements

We thank Sheila Luna, Christian Brunner and Jennifer Wilson for the artwork, and all of our collaborators who contributed thoughts and experiments. At the same time, we apologize to all authors whose work we could not cite owing to space limitations.
Correspondence and requests for materials should be addressed to A.G. or V.V.

Available online at www.sciencedirect.com
ScienceDirect
Mechanical Systems and Signal Processing 21 (2007) 2248–2266
Mechanical Systems and Signal Processing
www.elsevier.com/locate/jnlabr/ymssp

A segmental hidden semi-Markov model (HSMM)-based diagnostics and prognostics framework and methodology

Ming Dong[a,*], David He[b]

[a]Department of Industrial Engineering and Management, School of Mechanical Engineering (Room 618), Shanghai Jiao Tong University, 800 Dongchuan Road, Min-Hang District, Shanghai 200240, PR China
[b]Department of Mechanical and Industrial Engineering, The University of Illinois at Chicago, Chicago, IL 60607, USA

Received 30 April 2006; received in revised form 21 August 2006; accepted 4 October 2006
Available online 30 November 2006

Abstract

Diagnostics and prognostics are two important aspects in a condition-based maintenance (CBM) program. However, these two tasks are often separately performed. For example, data might be collected and analysed separately for diagnosis and prognosis. This practice increases the cost and reduces the efficiency of CBM and may affect the accuracy of the diagnostic and prognostic results.

In this paper, a statistical modelling methodology for performing both diagnosis and prognosis in a unified framework is presented. The methodology is developed based on segmental hidden semi-Markov models (HSMMs). An HSMM is a hidden Markov model (HMM) with temporal structures. Unlike HMM, an HSMM does not follow the unrealistic Markov chain assumption and therefore provides more powerful modelling and analysis capability for real problems. In addition, an HSMM allows modelling the time duration of the hidden states and therefore is capable of prognosis. To facilitate the computation in the proposed HSMM-based diagnostics and prognostics, new forward–backward variables are defined and a modified forward–backward algorithm is developed. The existing state duration estimation methods are inefficient because they require a huge storage and computational load. Therefore, a new approach is proposed for training HSMMs in which state duration probabilities are estimated on the lattice (or trellis) of observations and states. The model parameters are estimated through the modified forward–backward training algorithm. The estimated state duration probability distributions combined with state-changing point detection can be used to predict the useful remaining life of a system.

The evaluation of the proposed methodology was carried out through a real world application: health monitoring of hydraulic pumps. In the tests, the recognition rates for all states are greater than 96%. For each individual pump, the recognition rate is increased by 29.3% in comparison with HMMs. Because of the temporal structures, the same HSMMs can be used to predict the remaining-useful-life (RUL) of the pumps.

Keywords: Hidden semi-Markov model; Diagnostics; Prognostics; Integrated framework; State duration modelling; State-changing point

*Corresponding author. Tel.: +86 21 34206101.
E-mail address: mdong@sjtu.edu.cn (M. Dong).

0888-3270/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ymssp.2006.10.001

Nomenclature

N	number of states in the model
M	number of distinct observations for each state
$\{1, 2,\ldots,N\}$	N states in the model
s_t	state at time t
$V = \{v_1, v_2,\ldots, v_M\}$	observation symbols
$A = \{a_{ij}\}$	state transition probability distribution
$B = \{b_i(k)\}$	observation probability distribution in state i
$\pi = \{\pi_i\}$	initial state distribution
λ	an HMM model
$P_i(d)$	probability of the event of staying in i for exactly d time units
$O = o_1o_2\ldots o_T$	observation sequence
L	number of health states (or macro-states) in the model
$\{h_0, h_1,\ldots, h_{L-1}\}$	health states (or macro-states) in the model
d_l	duration staying at a health state h_l
T	life time of a component, $T = \sum_{l=0}^{L-1} d_l$
$\alpha_t(i)$	forward variable: probability of generating $o_1o_2\ldots o_t$ and ending in state i
D	maximum duration within any state
$b_j(O_{t-d+1}^t)$	joint density of d consecutive observations $(o_{t-d+1}o_{t-d+2}\ldots o_t)$
$\beta_t(i)$	backward variable: probability of generating $o_{t+1}o_{t+2}\ldots o_T$ and ending in state i
$\alpha_{t,t'}(i,j)$	probability of the partial observation sequence $o_1o_2\ldots o_{t'}$, and state i at time t and state j at time t' $(t' = t+d)$
$\phi_{t,t'}(i,j)$	probability of the system being in state i for d time units and then moving to the next state j
$\xi_{t,t'}(i,j)$	probability of the system being in state i for d $(t' = t+d)$ time units and then moving to the next state j, given the observation sequence $o_1o_2\ldots o_T$
$\bar{\pi}_i$	re-estimated initial state distribution
$\bar{a}_{ij}$	re-estimated state transition probability
$\bar{p}_i(d)$	re-estimated macro-state duration distribution
$\bar{b}_i(k)$	re-estimated segmental observation distribution
$\mu(i)$	mean of duration probability of macro-state i
$\sigma^2(i)$	variance of duration probability of macro-state i
η	Gaussian distribution with mean vector μ_{jm} and covariance matrix U_{jm} for the mth mixture component in the state j
M_j	number of Gaussian component η in state j
$c_{jm} = P(M_m = m\|s_t = j)$	conditional weight for the mth mixture component in the state j
$\hat{t}_{lc}$	weighted change-point time
t_{lc}	"true" change-point time
$D(h_l)$	state duration of h_l, which equals $\mu(h_l) + \rho\sigma^2(h_l)$
ρ	coefficient used for calculating $D(h_l)$, which equals $(T - \sum_{l=0}^{L-1}\mu(h_l))/\sum_{l=0}^{L-1}\sigma^2(h_l)$

1. Introduction

Diagnostics and prognostics are processes of assessment of a system's health. Diagnostics is an assessment about the current (and past) health of a system based on observed symptoms. It deals with fault detection, isolation and identification when fault occurs. Prognostics is an assessment of the future health, it is a task to determine whether a fault is impending and estimate how soon and how likely a fault will occur.

A prerequisite to effective wide-spread deployment of condition-based maintenance (CBM) practices is effective diagnostics and prognostics. CBM increases system efficiency and availability through elimination of

unnecessary maintenance. The economic ramifications of CBM are many folds since it affects labor requirements, replacement part costs, and the logistics of scheduling routine maintenance [1]. However, diagnostics and prognostics are often separately performed. For example, data might be collected and analysed separately for diagnosis and prognosis. Therefore, there is a need for an integrated framework in which both diagnostics and prognostics can be performed.

Due to the nature of the observed data and the available knowledge, the diagnostic and prognostic methods are often a combination of statistical inference and machine learning methods. A probabilistic approach called hidden Markov model (HMM) has been quite effective in some applications such as speech processing and medical diagnostics. Although the statistical methods of hidden Markov modelling were initially introduced and studied in the late 1960s and early 1970s, they have become increasingly popular recently. There are two major reasons for this. First, the models have very rich mathematical structure and can form the solid theoretical foundation for use. Second, the models have many successful applications in practice [2]. An added benefit of employing HMMs is the ease of model interpretation in comparison with pure "black-box" modelling methods such as artificial neural networks that are often employed in advanced diagnostic models [3].

An inherent limitation of HMM technology is that its state duration follows an exponential distribution. In other words, HMM does not provide adequate representation of temporal structure. Researchers have proposed a number of techniques to address these limitations. Ljolie and Levinson [4] use continuous variable duration HMM in the speech recognition. They compared the explicit duration model with the standard HMM. The results show that the absence of a correct duration model increases the error rate by 50%. The experimental evidence demonstrates that explicit duration models, even if only specifying the longest and the shortest duration allowed for a speech segment, can be beneficial to the recogniser performance [4–6]. As indicated by Chen et al. [7,8], the motivation for the variable state duration HMM-based handwritten word recognition (HWR) system is: because of the inherent ambiguity related to the segmentation process in handwritten words, it is a practical idea to use the variable duration model for the states in a HMM-based HWR system.

There has been some use of HMMs by the diagnostics community. Recently, some researchers apply HMMs in the area of diagnostics in machining processes [9–15]. However, in their applications, only ordinary HMM techniques are adopted. Therefore, the inherent limitation within HMMs as indicated above still exists in their models.

Literature on prognostic methods is extremely limited but the concept has been gaining importance in recent years. Unlike numerous methods available for diagnostics, prognostics is still in its infancy, and literature is yet to present a working model for effective prognostics [3]. Bunks et al. [1], and Baruah and Chinnam [3] first point out that HMM-based models could be applied in the area of prognostics in machining processes. However, only standard HMM-based approaches are proposed in their studies. The principle of HMM-based-prognostics in [3] is as follows: first, build and train N HMMs for all component health states. Between N trained HMMs, the authors assume that the estimated vectors of state transition times follow some multivariate distribution. Once the distribution is assessed, the conditional probability distribution of a distinct state transition given the previous state transition points can be estimated. In diagnostics of machining processes, tool wear is a time-related process. In prognostics of components, the objective is to predict the progression of a fault condition to component failure and estimate the remaining-useful-life (RUL) of the component. Component aging process is the critical point in this issue. Therefore, it is natural to use explicit state duration models.

In the proposed framework, for each health state of components, an hidden semi-Markov model (HSMM) is built and trained. Here, each health state of a component corresponds to a segment of the HSMM. These trained HSMMs can be used for the classification of a component failure mechanism given an observation sequence in diagnostics. For prognostics, another HSMM is used to model a component's life cycle. After training, the duration time in each health state can be estimated. From the estimated duration time, the proposed macro-state-based prognostic approach can be used to predict the remaining useful time for a component. The probability distribution for state transition times in [3] is estimated from the estimation of "state transition time instants". While in our model, the macro-state durations are estimated directly from the training data. Compared to the approach given in [3], our approach provides a unified HSMM-based

framework for both diagnostics and prognostics. The major drawback of HSMMs is that the computational complexity may increase for the inference procedures and parameter estimations. In this regard, duration modelling by parametric probability distributions has been used to alleviate the computational burden.

This paper is organised as follows. In Section 2, an HMM-based bearing diagnosis is provided as an introductory example and the basic concepts of HMMs are introduced. Section 3 presents the segmental HSMM-based modelling framework for diagnosis and prognosis. Section 4 provides inference and learning mechanisms for segmental HSMMs. Section 5 gives the HSMM-based diagnosis procedure. The HSMM-based prognosis algorithm is provided in Section 6. Case study for hydraulic pump health monitoring is given in Section 7. Finally, conclusions are drawn in Section 8.

2. Theoretical background

2.1. Description of fault diagnostic process using HMMs

The failure mechanisms of mechanical systems usually involve several degraded health states. For example, a small change in a bearing's alignment could cause a small nick in the bearing, which over time could cause scratches in the bearing race, which could then cause additional nicks, which could lead to complete bearing failure. This process can be ideally described by a mathematical model known as HMM since it can be used to estimate the unobservable health states using observable sensor signals. The word "hidden" means the HMM states are hidden from direct observations. In other words, the HMM states manifest themselves via some probabilistic behaviour. A diagram depicting the state transitions of the bearing from normal to the failure is shown in Fig. 1. The arrows represent transition paths from one state to another. For each path in HMM, there is usually a number denoting the transition probability associated with it. From Fig. 1, it can be seen that HMM is very suitable for describing the failure mechanism of the bearing example. One may argue that there may be infinite number of failure mechanisms in a bearing failure. This may be true. However, one should also note that the signals coming out of each step of failure process may have unique characteristics. For example, an experienced mechanics can distinguish a good transmission from a bad transmission by listening to the acoustic emissions of the transmission. HMM can exactly capture the characteristics of each stage of the failure process, which is the basis of using HMM for failure diagnosis and prognosis [16].

Vibration signals are preprocessed before extracting features that related to the specific fault. Preprocessing can be carried out by amplitude demodulating the signal. Amplitude demodulating provides a mechanism for effectively extracting the rolling element fault frequencies from extraneous noise that is present in the signal [17]. The preprocessed vibration signal for a bearing of known condition (or health state) is coded into a feature matrix. This feature matrix is then used to train an HMM that represents the specific bearing condition (or health state).

For the purpose of detecting the presence of a fault, it is sufficient to train a single HMM for the normal operating condition. Multiple vibration data collected from a normal bearing under various load conditions can be used to train an HMM. Presence of a fault can be detected in a bearing given the HMM for the normal condition. Given the feature matrix obtained from the bearing, the probability of the HMM for the normal condition is calculated. If the probability is above a pre-determined threshold, then there is no fault present in the bearing. The bearing has a fault, otherwise. This is summarised in Fig. 2.

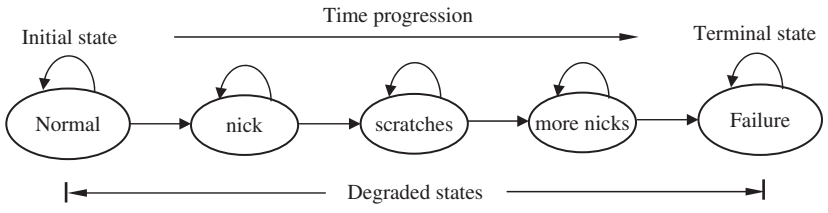


Fig. 1. An HMM scheme describing the failure mechanism of a bearing.

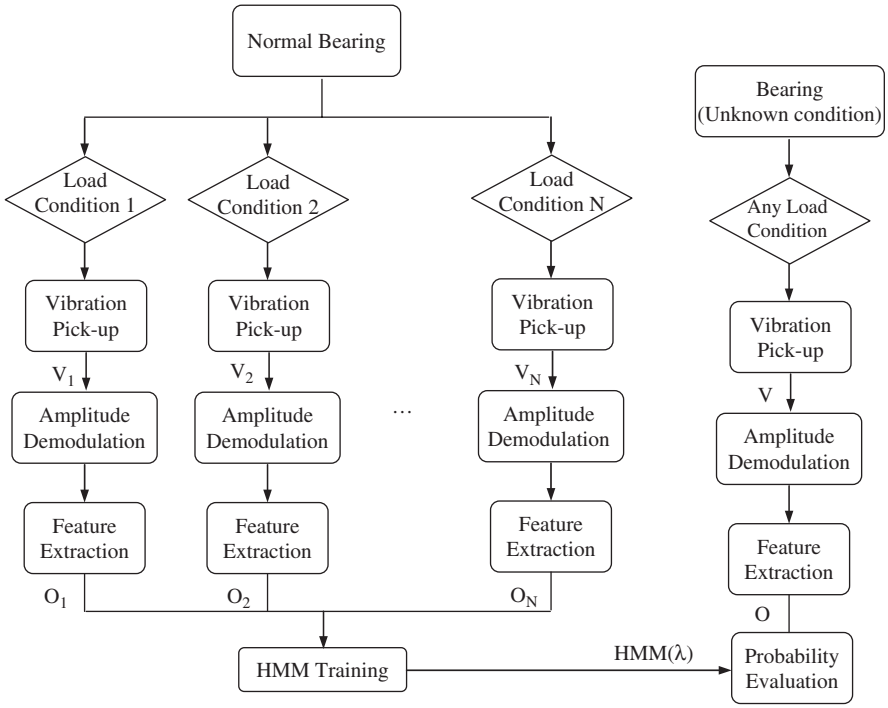


Fig. 2. HMM-based bearing fault detection.

2.2. Elements of an HMM

A Markov chain is a sequence of events, usually called states, the probability of each of which is dependent only on the event immediately preceding it. An HMM represents stochastic sequences as Markov chains where the states are not directly observed, but are associated with a probability function. The generation of a random sequence is then the result of a random walk in the chain and of an observation (also called an emission) at each visit of a state.

An HMM has the following elements [2]:

(1) N, the number of states in the model. Although the states are hidden, there is often some physical signal attached to the states of the model. We denote the individual states as $\{1, 2, \ldots, N\}$, and the state at time t as s_t.

(2) M, the number of distinct observations for each state. The observation symbols correspond to the physical output of the system being modelled. The individual observation symbols are denoted as $V = \{v_1, v_2, \ldots, v_M\}$.

(3) The state transition probability distribution $A = \{a_{ij}\}$, where

$$a_{ij} = P[s_{t+1} = j | s_t = i], \quad 1 \leqslant i, j \leqslant N.$$

(4) The observation probability distribution in state i, $B = \{b_i(k)\}$, where

$$b_i(k) = P[v_k | s_t = i], \quad 1 \leqslant i \leqslant N, 1 \leqslant k \leqslant M.$$

(5) The initial state distribution $\pi = \{\pi_i\}$ where

$$\pi_i = P[s_1 = i], \quad 1 \leqslant i \leqslant N.$$

It can be seen that a complete HMM requires the specifications of N, M, A, B, and π. For convenience, a compact notation is often used in the literature to indicate the complete parameter set of the model: $\lambda = (\pi, A, B)$.

2.3. Durational measure of standard HMMs

The durational behaviour of an HMM is usually characterised by a durational pdf $P(d)$. For a single state i, the value $P(d)$ is the probability of the event of staying in i for exactly d time units. This event is in fact the joint event of taking the self-loop for $(d-1)$ times and taking the out-going transition (with probability $1-a_{ii}$) just once. Given the Markovian assumption, and from probability theory, $P(d)$ is simply the product of all the d probabilities:

$$P_i(d) = a_{ii}^{d-1}(1 - a_{ii}). \tag{1}$$

Here, $P_i(d)$ denotes the probability of staying in state i for exactly d time steps, and a_{ii} is the self-loop probability of state i. It can be seen that this is a geometrically decaying function of d. It has been argued that this is a source of inaccurate duration modelling with the HMMs since most real-life applications would not obey this function [18].

2.4. The three basic problems for HMMs

In real applications, there are three basic problems associated with HMMs.

(1) *Evaluation* (also called *Classification*): Given the observation sequence $O = o_1 o_2 \ldots o_T$, and a HMM λ, what is the probability of the observation sequence given the model, i.e., $P(O|\lambda)$.
(2) *Decoding* (also called *Recognition*): Given the observation sequence $O = o_1 o_2 \ldots o_T$, and an HMM λ, what sequence of hidden states $S = s_1 s_2 \ldots s_T$ most probably generates the given sequence of observations.
(3) *Learning* (also called *Training*): How do we adjust the model parameters $\lambda = (A, B, \pi)$ to maximise $P(O|\lambda)$?

Different algorithms have been developed for above three problems. The most straightforward way of solving the evaluation problem is through enumerating every possible state sequence of length T (the number of observations). However, the computation burden for this exhaustive enumeration is prohibitively high. Fortunately, a more efficient algorithm that is based on dynamic programming exists. This algorithm is called forward–backward procedure [19]. The goal for decoding problem is to find the optimal state sequence associated with the given observation sequence. The most widely used optimality criterion is to find the single best state sequence (path), i.e., to maximise $P(S|O,\lambda)$ that is equivalent to maximising $P(S, O|\lambda)$. A formal technique for finding this single best state sequence exists, based on dynamic programming methods, and is called Viterbi algorithm [20]. For learning problem, there is no known way to obtain analytical solution. However, we can adjust the model parameters $\lambda = (A, B, \pi)$ such that $P(O|\lambda)$ is locally maximised using an iterative procedure such as the Baum–Welch method (or equivalently the Expectation-Maximisation algorithm) [21].

3. Segmental HSMM-based modelling framework for diagnostics and prognostics

3.1. Macro- and micro-states

For a component, it usually evolves through several distinct health-status prior to reaching failure. For example, mechanics of drilling processes suggest that a typical drill-bit may go through four health-states: good, medium, bad, and worst. In general, for a component, we can identify L distinct sequential states for a failure mechanism. That is, determination of health status of a component: no-defect (i.e., health state 0, denoted by h_0), level-1 defect (denoted by h_1), level-2 defect (denoted by h_2), …, level-$(L-1)$ defect (denoted by h_{L-1}). Here, the level-$(L-1)$ defect means failure. Let d_l be the duration staying at the health state h_l and T be the life time of the component. Then, $T = \sum_{l=0}^{L-1} d_l$.

Unlike a state in a standard HMM, a state in a segmental semi-Markov model generates a segment of observations, as opposed to a single observation in the HMM. In this study, the states in a segmental semi-Markov model are called *macro-states* (i.e., segments). Each macro-state consists of several single states, which are called *micro-states*. Suppose that a macro-state sequence has L segments, and let q_l be the time index of the end-point of the lth segment ($0 \leqslant l \leqslant L-1$). The segments are as follows:

Time units	$\underline{1,\ldots,q_0}$	$\underline{q_0+1,\ldots,q_1}$	$\underline{q_{L-2}+1,\ldots,q_{L-1}}$
Observations	$\underline{o_1,\ldots,o_{q_0}}$	$\underline{o_{q_0+1},\ldots,o_{q_1}}$	$\underline{o_{q_{L-2}+1},\ldots,o_{q_{L-1}}}$
Micro-states	$\underline{s_1,\ldots,s_{q_0}}$	$\underline{s_{q_0+1},\ldots,s_{q_1}}$	$\underline{s_{q_{L-2}+1},\ldots,s_{q_{L-1}}}$
Durations	$d_0 = q_1$	$d_1 = q_1 - q_0$	$d_{L-1} = q_{L-1} - q_{L-2}$
Macro-states	h_0	h_1	h_{L-1}
Segments	0	1	$L-1$

For the lth macro-state, the observations are $o_{q_{l-1}+1},\ldots,o_{q_l}$, and they have the same micro-state label:

$$s_{q_{l-1}+1} = s_{q_{l-1}+2} = \cdots = s_{q_l} \equiv h_l.$$

The proposed segmental HSMM-based modelling framework for component diagnostics and prognostics is described in Fig. 3.

3.2. Model structure

3.2.1. Model description

Let s_t be the hidden state at time t and O be the observation sequence. Characterisation of an HSMM is through its parameters. The parameters for an HSMM are: the initial state distribution (denoted by π), the transition model (denoted by A), state duration distribution (denoted by D), and the observation model (denoted by B). Thus, an HSMM can be written as $\lambda = (\pi, A, D, B)$.

3.2.2. State transitions

In the segmental HSMM, there are N states, and the transitions between the states are according to the transition matrix A, i.e., $P(i \to j) = a_{ij}$. Similar to standard HMMs, we assume that the state s_0 at time $t = 0$ is a special state "START". We denote this initial state distribution as π.

Although the macro-state transition $s_{q_{l-1}} \to s_{q_l}$ is Markov

$$P(s_{q_l} = j | s_{q_{l-1}} = i) = a_{ij}$$

the micro-state transition $s_{t-1} \to s_t$ is usually not Markov. This is the reason why the model is called "semi-Markov" [22]. That is, in the HSMM case, the conditional independence between the past and the future is only ensured when the process moves from one state to another distinct state.

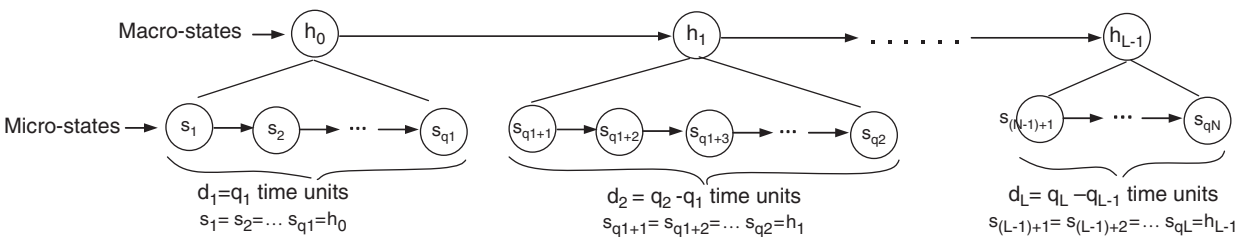


Fig. 3. Segmental HSMM-based modelling framework for component diagnostics and prognostics.

3.2.3. Segmental observation distributions

Another extension in segmental HSMM from the HMM is the segmental observation distribution. The observations $o_{(t_1,t_2]}$ in a segment with state i and duration d are produced by

$$P(o_{(t_1,t_2]}|i,d), \tag{2}$$

where $d = t_2 - t_1$.

4. Inference and leaning mechanisms for segmental HSMM-based diagnostics and prognostics framework

4.1. Inference procedures

Similar to HMMs, HSMMs also have three basic problems to deal with, i.e., evaluation, recognition and training problems. To facilitate the computation in the proposed HSMM-based diagnostics and prognostics framework, in the following, new forward–backward variables are defined and modified forward–backward algorithm is developed.

A dynamic programming scheme is employed for the efficient computation of the inference procedures. To implement the inference procedures, a *forward variable* $\alpha_t(i)$ is defined as the probability of generating $o_1 o_2 \ldots o_t$ and ending in state i:

$$\alpha_t(i) = P(o_1 o_2 \ldots o_t, i \text{ ends at } t|\lambda) = \sum_{i=1}^{N} \sum_{d=1}^{\min(D,t)} \alpha_{t-d}(i) a_{ij} P(d|j) b_j(O_{t-d+1}^{t}), \tag{3}$$

where D is the maximum duration within any state. $b_j(O_{t-d+1}^{t})$ is the joint density of d consecutive observations $(o_{t-d+1} o_{t-d+2} \ldots o_t)$.

It can be seen that the probability of O given the model λ can be written as

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i). \tag{4}$$

A modified Forward–Backward algorithm for HSMMs is given in Appendix A. The parameter re-estimations for HSMM-based diagnosis and prognosis can be found in Appendix B.

4.2. Training of macro-state duration models using parametric probability distributions

In this study, state duration densities are modelled by single Gaussian distribution estimated from training data. The existing state duration estimation method is through the simultaneous training HSMMs and their state duration densities. However, these techniques are inefficient because it requires huge storage and computational load. Therefore, we adopt a new approach for training state duration models. In this approach, state duration probabilities are estimated on the lattice (or trellis) of observations and states which is obtained in the HSMM training stage. For detailed parameter re-estimation formulae of macro-state durations, see Appendix C.

4.3. Continuous observation densities in HSMMs

Although the vector quantisation (VQ) can be used to quantise signals via codebook, there might be serious degradation associated with such quantisation [2]. Hence it would be advantageous to use the HSMMs with continuous observation densities. In this research, mixture Gaussian distribution is used. The most general representation of the pdf is a finite mixture of the form:

$$b_j(O) = \sum_{m=1}^{M_j} c_{jm} \eta[O, \mu_{jm}, U_{jm}] = \sum_{m=1}^{M_j} P(M_m = m|s_t = j) \eta[O, \mu_{jm}, U_{jm}], \quad 1 \leqslant j \leqslant N, \tag{5}$$

where η represents a Gaussian distribution with mean vector μ_{jm} and covariance matrix U_{jm} for the mth mixture component in the state j, O is the vector being modelled, M_j is the number of Gaussian component η

in state j, $c_{jm} = P(M_m = m|s_t = j)$ is the conditional weight for the mth mixture component in the state j. The mixture gains c_{jm} satisfy the following stochastic constraint:

$$\sum_{m=1}^{M_j} c_{jm} = 1, \quad 1 \leqslant j \leqslant N,$$
$$c_{jm} \geqslant 0, \quad 1 \leqslant j \leqslant N, \quad 1 \leqslant m \leqslant M_j, \tag{6}$$

so that the pdf is properly normalised, i.e.,

$$\int_{-\infty}^{\infty} b_j(x)\,\mathrm{d}x = 1, \quad 1 \leqslant j \leqslant N. \tag{7}$$

As pointed out in [2], the pdf of Eq. (5) can be used to approximate, arbitrarily closely, any finite continuous density function of practical importance. Hence, it can be applied to a wide range of problems.

5. Segmental HSMM-based diagnosis

5.1. HSMM training

For diagnostics, the goal is to develop trained HSMMs to recognise L different health states of a component for a given failure mode. That is, the task is to develop a diagnostics model for classifying the health states of a component. Therefore, for the diagnosis of the fault, it is necessary that a separate HSMM be trained for all possible fault types in addition to the HSMM for normal condition.

Given these L groups of observation sequences, L different HSMMs (i.e., HSMM_0, HSMM_1,…, HSMM_{L-1}) are modelled for characterisation of each group, which corresponds to a health state.

5.2. Classification using HSMMs

The procedure for classification of a component failure mechanism given an observation sequence is: each of the N trained HSMMs is presented with the same sequence. According to the value of highest log-likelihood, the sequence can be classified. The training phase, classification phase and prediction phase of an HSMM-based fault diagnosis and prognosis scheme are illustrated in Fig. 4.

6. Segmental HSMM-based prognosis

6.1. Health-state change point detection

The health-state change point is defined as the point at which the system changes from health state h_l to health state h_{l+1}. Through the health-state change point detection, we can estimate the time from the system's current condition to the health-state change point. As the change-point corresponds to the switching from health state h_l to health state h_{l+1} in the model, the detection of change-point is straightforward: we run the Viterbi algorithm on-line as new data points $o_1 \ldots\ o_t \ldots$ are coming in. If $s_t = h_{l+1}$ in the most likely state sequence, t will be the change-point. For a given sequence of observed data $O = o_1 o_2 \ldots o_T$, and a 2-health-state system, there are, in theory, $T-1$ possible state sequences are as shown in Fig. 5.

Each state sequence provides an estimate of the location of the change-point. For example, in $\mathbf{s}^{(t)}$ the location of change-point is t.

Since the weighted change-point time can be estimated as

$$\hat{t}_{lc} = \sum_{t=2}^{T} t \times P(s^{(t)}|O). \tag{8}$$

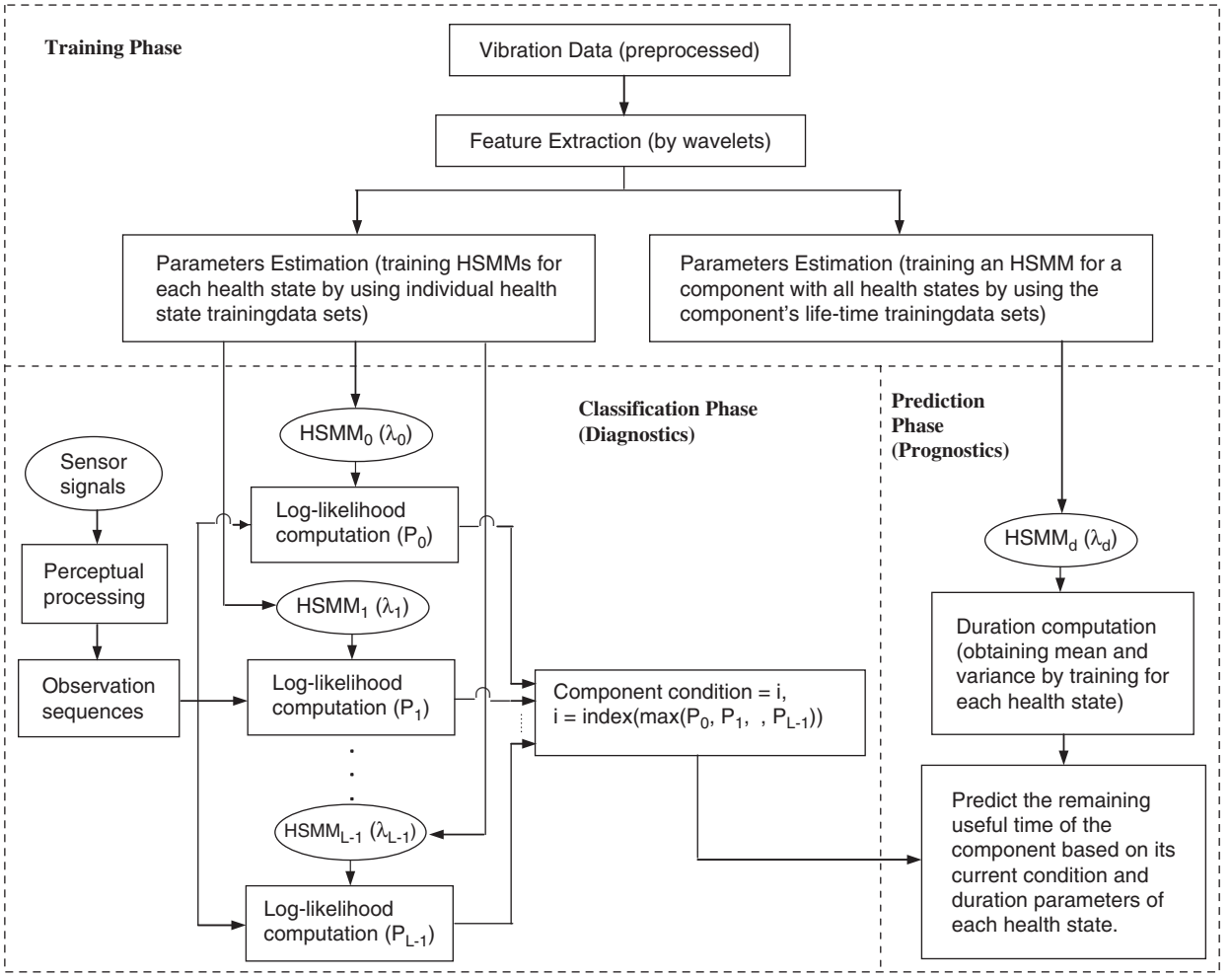


Fig. 4. HSMM-based fault diagnosis and prognosis scheme.

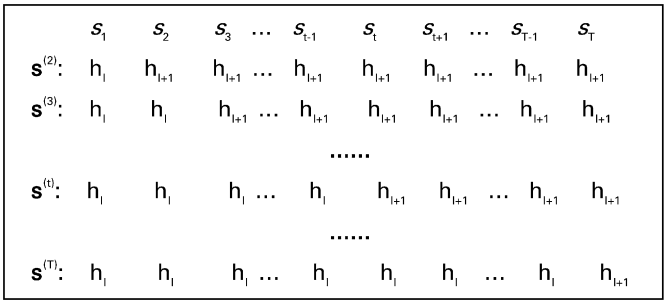


Fig. 5. Possible state sequences for 2-health-state system.

Note that

$$P(s_{t-1}=l|O)=\sum_{t'=t}^{T}P(s^{(t')}|O)=P(s^{(t)}|O)+\sum_{t'=t+1}^{T}P(s^{(t')}|O)=P(s^{(t)}|O)+P(s_t=l|O). \tag{9}$$

Therefore,

$$\hat{t}_{lc} = \sum_{t=2}^{T} (t \times [P(s_{t-1} = l|O) - P(s_t = l|O)]). \tag{10}$$

This weighted average minimises the penalty function $\sum_{t_{lc}} P(t_{lc}|O)(t_{lc} - \hat{t}_{lc})^2$, where t_{lc} is the "true" change-point time. The above estimator can be efficiently computed by the modified Forward–Backward algorithm.

6.2. Macro-state duration model-based prognostics

The objective of prognostics is to predict the progression of a fault condition to component failure and estimate the RUL of the component. In the following, we propose an approach that is based on macro-state duration models.

The framework for macro-state duration model-based prognostics is given in Fig. 6. Since each macro-state duration density $P(d_n|h_l)$ is modelled by a single Gaussian distribution, state durations, which maximise $\log P(S|\lambda, T) = \sum_{l=0}^{L-1} \log P(d_n|h_l)$ under the constraint $T = \sum_{l=0}^{L-1} D(h_l)$, are given by

$$D(h_l) = \mu(h_l) + \rho\sigma^2(h_l), \tag{11}$$

$$\rho = \left(T - \sum_{l=0}^{L-1} \mu(h_l)\right) \Big/ \sum_{l=0}^{L-1} \sigma^2(h_l). \tag{12}$$

6.3. Prognostics procedure

The macro-state duration model-based component prognostics procedure is given as follows:

Step 1: From the HSMM training procedure (i.e., parameter estimation), we can obtain the state transition probability for HSMM.

Step 2: Through the HSMM parameter estimation, the duration pdf for each macro-state can be obtained. Therefore, the duration mean and variance can be calculated.

Step 3: By classification, identify the current health status of the component.

Step 4: The RUL of the system can be computed by the following backward recursive equations (suppose that the system currently stays at health state l, RUL_l indicates the RUL starting from state l):

At state $L-2$:

$$\mathrm{RUL}_{L-2} = a_{L-2,L-2}[D(h_{L-2}) + D(h_{L-1})] + a_{L-2,L-1}[D(h_{L-1})].$$

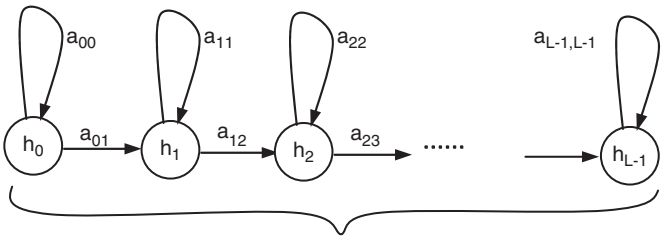


Fig. 6. Segmental HSMM for prognostics.

At state $L-3$:

$$\mathrm{RUL}_{L-3} = a_{L-3,L-3}[D(h_{L-3}) + \mathrm{RUL}_{L-2}] + a_{L-3,L-2}[\mathrm{RUL}_{L-2}]$$

……

At state l:

$$\mathrm{RUL}_l = \hat{t}_{lc} + a_{l,l+1}[\mathrm{RUL}_{l+1}]. \tag{13}$$

7. Case studies

7.1. Diagnostics for pumps

To evaluate the performance of the developed HSMM method for machine health diagnosis and prognosis, it was tested using data from a real hydraulic pump health monitoring application case study.

In this case study, long-term wear test experiments were conducted at a research laboratory facility. In the test experiments, three pumps (pump 6, pump 24 and pump 82) were worn to various percent decreases in flow by running them using oil containing dust. Each pump experienced four states: baseline (normal state), contamination 1 (5 mg of 20-μm dust injected into the oil reservoir), contamination 2 (10 mg of 20-μm dust injected into the oil reservoir), and contamination 3 (15 mg of 20-μm dust injected into the oil reservoir). The contamination stages in this hydraulic pump wear test case study correspond to different stages of flow loss in the pumps. As flow rate of a pump clearly indicates the heath state of a pump, therefore, the contamination stages corresponding to different degrees of flow loss in a pump were defined as the health states of the pump in the pump wear test.

Vibration signals were collected from a pump accelerometer that was positioned parallel to the axis of swash plate swivel axis. Fig. 7 shows the schematic diagram of the experimental set-up. The pump used for testing in the experiments was a Back Hoe Loader: a 74 cm^3/rev Variable Displacement Pump.

The data was collected at a sample rate of 60 kHz with anti-aliasing filters from accelerometers designed to have a usable range of 10 kHz. These signals were processed using wavelet packet with Daubechies wavelet 10 (db10) and five decomposition levels as the db10 wavelet with five decomposition levels provides the most effective way to capture the fault information in the pump vibration data [23,24]. The wavelet coefficients obtained by the wavelet packet decomposition were used as the inputs to the HMMs and HSMMs. In this test, we wanted to see how the HSMMs could classify the health conditions of the pumps in comparison with the HMMs. The number of data points used in training and testing for each condition is provided in Table 1.

The diagnosis results for three pumps are given in Tables 2–4, respectively.

The classification rate can be calculated as follows:

For pump 6, the classification rate based on HMMs is: $4/7 = 57\%$. The classification rate based on HSMMs is: $7/7 = 100\%$.

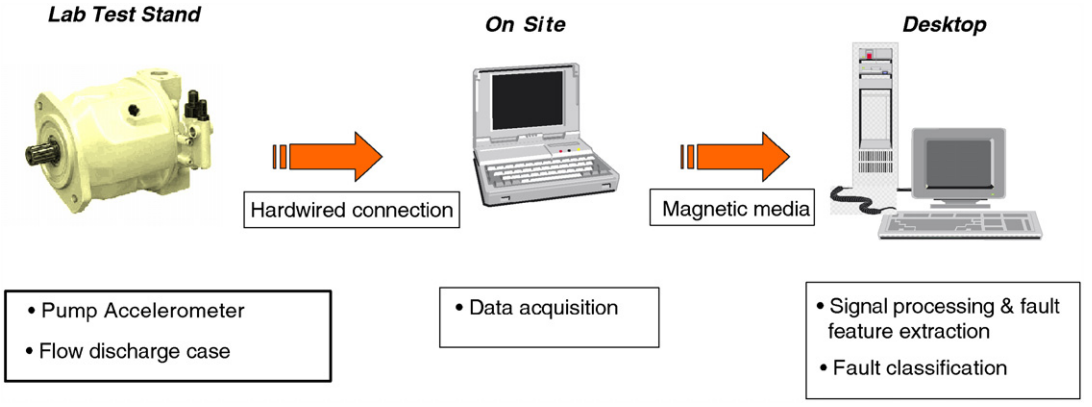


Fig. 7. Schematic diagram of the experimental set-up.

Table 1
Number of data points used for training and testing under different pump conditions

Pump no.	Pump conditions							
	Baseline		Contamination 1		Contamination 2		Contamination 3	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Pump 6	12	2	8	1	11	2	8	2
Pump 24	8	2	8	2	10	2	8	2
Pump 82	8	1	8	2	8	2	9	2

Table 2
Diagnosis results of pump 6 based on HMM and HSMM (× indicates a wrong classification)

Log-likelihood value	Baseline Test run 1	Baseline Test run 2	Contamination 1 Test run 1	Contamination 2 Test run 1	Contamination 2 Test run 2	Contamination 3 Test run 1	Contamination3 Test run 2
HMM_B	−135.9974	−105.5929	-Inf	-Inf	-Inf	-Inf	-Inf
HMM_{C1}	-Inf	-Inf	−16.2375	-Inf	-Inf	-Inf	-Inf
HMM_{C2}	-Inf	-Inf	-Inf	-Inf (×)	-Inf (×)	-Inf	-Inf
HMM_{C3}	-Inf	-Inf	-Inf	-Inf	-Inf	−177.5376	-Inf (×)
$HSMM_B$	−0.2787	15.4053	-Inf	-Inf	-Inf	-Inf	-Inf
$HSMM_{C1}$	-Inf	-Inf	21.9241	-Inf	-Inf	-Inf	-Inf
$HSMM_{C2}$	-Inf	-Inf	−611.7341	−25.7057	8.8795	−215.1119	−204.3999
$HSMM_{C3}$	-Inf	-Inf	-Inf	−716.2021	−327.9671	16.3517	2.7867

For pump 24, the classification rate based on HMMs is: $6/8 = 75\%$. The classification rate based on HSMMs is: $8/8 = 100\%$.

For pump 82, the classification rate based on HMMs is: $7/7 = 100\%$. The classification rate based on HSMMs is: $7/7 = 100\%$.

The results verify that the new scheme is able to detect and diagnose faults with 100% accuracy. For individual pump's diagnostics, it can be seen that the correct recognition rate is increased by 29.3%, which shows the proposed model is superior to currently used HMM-based approach. In addition, experiments show that both HMM-based diagnosis and HSMM-based diagnosis have almost the same computational time. This means that HSMM-based method is efficient and could be used in the real applications with large data sets.

7.2. Prognostics for pumps

For prognostics, we use the life time training data from pump 6, pump 24 and pump 82. By training, an HSMM with four health states can be obtained. And, the mean and variance of the duration time in each state are also available through the training process. The results are given in Tables 5 and 6.

Based on above information, the mean value of the RUL of a pump can be calculated as follows (in terms of Eq. (13) and suppose that the component currently stays at state "Contamination1"):

$$RUL_{mean} = 28.0829.$$

Similarly, the variance of the RUL of a pump can be obtained as follows:

$$RUL_{variance} = 1.7846.$$

That is, if the component is currently at state "Contamination1", then its expected RUL is 28.0829 time units with confidence interval 1.7846 time units.

Table 3
Diagnosis results of pump 24 based on HMM and HSMM

Log-likelihood value	Baseline Test run 1	Baseline Test run 2	Contamination 1 Test run 1	Contamination 1 Test run 2	Contamination 2 Test run 1	Contamination 2 Test run 2	Contamination3 Test run 1	Contamination3 Test run 2
HMM_B	35.2932	35.5924	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
HMM_{C1}	-Inf	-Inf	−35.9807	−21.1165	-Inf	-Inf	-Inf	-Inf
HMM_{C2}	-Inf	-Inf	-Inf	-Inf	9.5903	−80.6109	-Inf	-Inf
HMM_{C3}	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf (×)	-Inf (×)
$HSMM_B$	29.9232	29.8745	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
$HSMM_{C1}$	-Inf	-Inf	11.4684	18.7377	−577.3743	−589.9719	-Inf	-Inf
$HSMM_{C2}$	-Inf	-Inf	-Inf	-Inf	19.0375	18.1357	-Inf	-Inf
$HSMM_{C3}$	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	−73.7291	11.6900

Table 4
Diagnosis results of pump 82 based on HMM and HSMM

Log-likelihood value	Baseline Test run 1	Contamination 1 Test run 1	Contamination 1 Test run 2	Contamination 2 Test run 1	Contamination 2 Test run 2	Contamination 3 Test run 1	Contamination 3 Test run 2
HMM_B	33.2111	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
HMM_{C1}	-Inf	−73.3233	−4.1633	-Inf	-Inf	-Inf	-Inf
HMM_{C2}	-Inf	-Inf	-Inf	−217.4598	−4.7148	-Inf	-Inf
HMM_{C3}	-Inf	-Inf	-Inf	-Inf	-Inf	29.3313	11.2834
$HSMM_B$	27.4588	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
$HSMM_{C1}$	−584.5710	9.9819	16.0810	-Inf	−460.2582	-Inf	-Inf
$HSMM_{C2}$	-Inf	-Inf	-Inf	−11.4333	21.1464	-Inf	-Inf
$HSMM_{C3}$	-Inf	-Inf	-Inf	-Inf	-Inf	22.5104	18.2022

Table 5
Transition probability between four health states

States	Baseline	Contamination 1	Contamination 2	Contamination 3
Baseline	0.8913	0.0454	0.0633	0.0000
Contamination 1	0.0000	0.6399	0.3599	0.0003
Contamination 2	0.0000	0.0000	0.9167	0.0833
Contamination 3	0.0000	0.0000	0.0000	1.0000

Table 6
Mean and variance of duration time in four health states

States	Baseline	Contamination 1	Contamination 2	Contamination 3
Mean of duration	10.4549	9.7923	11.3375	10.4793
Variance of duration	1.9388	0.9792	1.2415	0.1880

8. Conclusions

In this paper, an integrated system for component diagnostics and prognostics is presented. Usually, diagnostics and prognostics are separately performed. The proposed segmental HSMM-based framework combines diagnostics and prognostics in an integrated manner. The health states of a component are modelled by state transition probability matrix and observation probability. The duration of each health segment is described by the state duration probability. As a whole, they are modelled as a hidden semi-Markov chain.

A modified Forward–Backward algorithm for segmental HSMMs is provided to estimate the parameters of HSMMs. To facilitate the computational procedure, new forward and backward variables are defined. And correspondingly, the re-estimation formulae based on new variables are derived. By incorporating the explicit temporal structure into the framework, the HSMM is enabled to predict the useful remaining life of components. For prognostics, macro-state duration model-based prediction procedure is provided.

The case studies show that HSMM-based diagnostics has a much better performance than HMM-based diagnostics. From our observations, there are two reasons that contribute the performance improvement of HSMM-based diagnostics than HMM: (1) in HMMs, the exponential distribution is implicitly used to approximate the durations associated with the states. While in HSMMs, the explicit probability distributions such as Gaussian distribution are adopted to model the state durations more accurately. (2) HMMs only carry the information obtained from the previous state. While in HSMMs, the decision-making is based on the information from previous d states. In other words, more information is used in HSMMs than HMMs.

Acknowledgments

The authors express their deepest gratitude to anonymous referees for their comments and suggestions on this research. The presentation of the paper has significantly improved with their inputs.

Appendix A. Modified Forward–Backward algorithm for segmental HSMMs

Similar to forward variable, the backward variable can be written as

$$\beta_t(i) = \sum_{j=1}^{N} \sum_{d=1}^{\min(D,t)} a_{ij} P(d|j) b_j(O_{t+1}^{t+d}) \beta_{t+d}(j). \tag{A.1}$$

In order to give re-estimation formulas for all variables of the HSMM, three more segment-featured forward–backward variables are defined

$$\alpha_{t,t'}(i,j) = P(o_1 o_2 \dots o_{t'}, t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j|\lambda), \tag{A.2}$$

$$\phi_{t,t'}(i,j) = \sum_{d=1}^{D} [P(d = t' - t|j) P(O_{t+1}^{t'}|t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j, \lambda)], \tag{A.3}$$

$$\xi_{t,t'}(i,j) = P(t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j|O_1^T, \lambda). \tag{A.4}$$

Here, $O_{t+1}^{t'} = o_{t+1} o_{t+2} \dots o_{t'}$ and $O_1^T = o_1 o_2 \dots o_T$.

$\phi_{t,t'}(i,j)$ is the probability of the system being in state i for d time units and then moving to the next state j. $\alpha_{t,t'}(i,j)$ can be described, in terms of $\phi_{t,t'}(i,j)$, as follows:

$$\begin{aligned}
\alpha_{t,t'}(i,j) &= P(o_1 o_2 \dots o_t o_{t+1} \dots o_{t'}, t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j|\lambda) \\
&= P(o_1 o_2 \dots o_t, t = q_n, s_t = i|\lambda) P(O_{t+1}^{t'}, t' = q_{n+1}, s_{t'} = j|O_1^t, t = q_n, s_t = i, \lambda) \\
&= \alpha_t(i) P(O_{t+1}^{t'}, t' = q_{n+1}, s_{t'} = j|t = q_n, s_t = i, \lambda) \\
&= \alpha_t(i) P(t' = q_{n+1}, s_{t'} = j|t = q_n, s_t = i, \lambda) P(O_{t+1}^{t'}|t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j, \lambda) \\
&= \alpha_t(i) a_{ij} \sum_{d=1}^{D} P(d = t' - t|j) P(O_{t+1}^{t'}|t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j, \lambda) \\
&= \alpha_t(i) a_{ij} \phi_{t,t'}(i,j).
\end{aligned} \tag{A.5}$$

The relationship between $\alpha_t(i)$ and $\alpha_{t,t'}(i,j)$ is given in the following:

$$\begin{aligned}
\alpha_{t'}(j) &= P(o_1 o_2 \dots o_{t'}, t' = q_{n+1}, s_{t'} = j|\lambda) \\
&= \sum_{i=1}^{N} \sum_{d=1}^{D} P(d = t' - t|j) P(o_1 o_2 \dots o_t o_{t+1} \dots o_{t'}, t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j|\lambda) \\
&= \sum_{i=1}^{N} \sum_{d=1}^{D} P(d = t' - t|j) \alpha_{t,t'}(i,j).
\end{aligned} \tag{A.6}$$

From the definitions of the forward–backward variables, we can derive $\xi_{t,t'}(i,j)$ as follows:

$$\begin{aligned}
\xi_{t,t'}(i,j) &= P(t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j|O_1^T, \lambda) \\
&= \frac{P(t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j, O_1^T|\lambda)}{P(O_1^T|\lambda)}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{P(t=q_n, s_t=i, t'=q_{n+1}, s_{t'}=j, O_1^t, O_{t+1}^{t'}, O_{t'+1}^T|\lambda)}{P(O_1^T|\lambda)}\\
&= \frac{P(t=q_n, s_t=i, O_1^t|\lambda)P(t'=q_{n+1}, s_{t'}=j, O_{t+1}^{t'}, O_{t'+1}^T|t=q_n, s_t=i, O_1^t, \lambda)}{P(O_1^T|\lambda)}\\
&= \frac{\alpha_t(i)P(t'=q_{n+1}, s_{t'}=j, O_{t+1}^{t'}, O_{t'+1}^T|t=q_n, s_t=i, O_1^t, \lambda)}{P(O_1^T|\lambda)}\\
&= \frac{\alpha_t(i)P(t'=q_{n+1}, s_{t'}=j, O_{t+1}^{t'}, O_{t'+1}^T|t=q_n, s_t=i, \lambda)}{P(O_1^T|\lambda)}\\
&= \frac{\alpha_t(i)P(t'=q_{n+1}, s_{t'}=j|t=q_n, s_t=i, \lambda)P(O_{t+1}^{t'}, O_{t'+1}^T|t=q_n, s_t=i, t'=q_{n+1}, s_{t'}=j)}{P(O_1^T|\lambda)}\\
&= \frac{\alpha_t(i)a_{ij}\sum_{d=1}^{D}P(d|j)P(O_{t+1}^{t'}|t=q_n, s_t=i, t'=q_{n+1}, s_{t'}=j)P(O_{t'+1}^T|t=q_n, s_t=i, t'=q_{n+1}, s_{t'}=j, O_{t+1}^{t'})}{P(O_1^T|\lambda)}\\
&= \frac{\alpha_t(i)a_{ij}\sum_{d=1}^{D}P(d|j)b_j(O_{t+1}^{t'})P(O_{t'+1}^T|t'=q_{n+1}, s_{t'}=j)}{P(O_1^T|\lambda)}\\
&= \frac{\alpha_t(i)a_{ij}\sum_{d=1}^{D}P(d|j)b_j(O_{t+1}^{t'})\beta_{t'}(j)}{P(O_1^T|s_0=\mathrm{START}, \lambda)}\\
&= \frac{\sum_{d=1}^{D}\alpha_t(i)a_{ij}\phi_{t,t'}(i,j)\beta_{t'}(j)}{\beta_0(i=S\mathrm{START})}.
\end{aligned}
\tag{A.7}
$$

The Forward–Backward algorithm computes the following probabilities:
Forward pass: The forward pass of the algorithm computes $\alpha_t(i)$, $\alpha_{t,t'}(i,j)$ and $\phi_{t,t'}(i,j)$.
Step 1: Initialisation ($t=0$)

$$
\alpha_{t=0}(i) = \begin{cases} 1 & \text{if } i=\mathrm{START},\\ 0 & \text{otherwise.}\end{cases}
$$

Step 2: Forward recursion ($t>0$). For $t=1,2,\ldots,T$; $1\leqslant i, j\leqslant N$, and $1\leqslant d\leqslant D$.

$$
\begin{aligned}
&\phi_{t,t'}(i,j) = \sum_{d=1}^{D}[P(d=t'-t|j)P(O_{t+1}^{t'}|t=q_n, s_t=i, t'=q_{n+1}, s_{t'}=j, \lambda)],\\
&\alpha_{t,t'}(i,j) = \alpha_t(i)a_{ij}\phi_{t,t'}(i,j),\\
&\alpha_{t'}(j) = \sum_{i=1}^{N}\sum_{d=1}^{D}P(d=t'-t|j)\alpha_{t,t'}(i,j).
\end{aligned}
$$

Backward pass: The backward pass computes $\beta_t(i)$ and $\xi_{t,t'}(i,j)$.
Step 1: Initialisation ($t=T$ and $1\leqslant i, j\leqslant N$)

$$\beta_T(i) = 1.$$

Step 2: Backward recursion ($t<T$). For $t=1,2,\ldots,T$; $1\leqslant i, j\leqslant N$; and $1\leqslant d\leqslant D$

$$
\beta_t(i) = \sum_{j=1}^{N}\sum_{d=1}^{\min(D,t)}a_{ij}P(d|j)b_j(O_{t+1}^{t+d})\beta_{t+d}(j) = \sum_{j=1}^{N}a_{ij}\phi_{t,t'}(i,j)\beta_{t'}(j),
\tag{A.8}
$$

$$
\xi_{t,t'}(i,j) = \sum_{d=1}^{D}\alpha_t(i)a_{ij}\phi_{t,t'}(i,j)\beta_{t'}(j)\Big/\beta_0(i=\mathrm{START}).
$$

Let D_l be the maximum duration for state l. The total computational complexity for the forward-backward algorithm is $O(N^2DT)$, where $D = \sum_{l=0}^{L-1} D_l$.

Appendix B. Parameter re-estimation for the segmental HSMM-based diagnostics and prognostics framework

B.1. Initial state distribution

The re-estimation formula for initial state distribution is the probability that state i was the first state, given O

$$\bar{\pi}_i = \frac{\pi_i[\sum_{d=1}^{D}\beta_d(i)P(d|i)b_j(O_1^d)]}{P(O|\lambda)}. \tag{B.1}$$

B.2. State transition probabilities

The re-estimation formula of state transition probabilities is the ratio of the expected number of transitions from state i to state j, to the expected number of transitions from state i

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T}\xi_{t,t'}(i,j)}{\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N}\xi_{t,t'}(i,j)}. \tag{B.2}$$

B.3. Macro-state duration distributions

The formula of state duration distributions is the ratio of the expected number of times state i occurred with duration d, to the expected number of times state i occurred with any duration

$$\bar{p}_i(d) = \frac{\sum_{t=1}^{T}\alpha_{t,t'}(i,j)P(d=t'-t|j)b_j(O_{t+1}^{t'})}{\sum_{d=1}^{D}\sum_{t=1}^{T}\alpha_{t,t'}(i,j)P(d=t'-t|j)b_j(O_{t+1}^{t'})}. \tag{B.3}$$

B.4. Segmental observation distributions

The re-estimation formula for segmental observation distributions is the expected number of times that observation $o_t = v_k$ occurred in state i, normalised by the expected number of times that any observation occurred in state i. Since $\alpha_t(i)$ accounts for the partial observation sequence $o_1o_2\ldots o_t$ and state i at t, while $\beta_t(i)$ accounts for the partial observation sequence $o_{t'}o_{t'+1}\ldots o_T$, given state i at t. The remainder of the observation sequence $o_to_{t+1}\ldots o_{t'}$ given state i at t and state j at t' is accounted by $P(O_{t+1}^{t'}|t=q_n, s_t=i, t'=q_{n+1}, s_{t'}=j)$. Therefore, the re-estimation of segmental observation distributions can be calculated as follows:

$$\bar{b}_i(k) = \frac{\sum_{\substack{t=1 \\ \text{s.t. } O_t=v_k}}^{T}\alpha_t(i)\left[\frac{\phi_{t,t'}(i,j)}{\sum_{d=1}^{D}P(d=t'-t|i)}\right]\beta_t(i)}{\sum_{t=1}^{T}\alpha_t(i)\left[\frac{\phi_{t,t'}(i,j)}{\sum_{d=1}^{D}P(d=t'-t|i)}\right]\beta_t(i)}. \tag{B.4}$$

Appendix C. Parameter re-estimation for macro-state duration

The mean $\mu(l)$ and the variance $\sigma^2(l)$ of duration probability of health-state l are determined by

$$\mu(l) = \frac{\sum_{q_{n-1}=1}^{T}\sum_{q_n=q_{n-1}}^{T}\frac{1}{\sqrt{2\pi}\sigma}\,\mathrm{e}^{-\{[(q_n-q_{n-1})-\mu]^2/2\sigma^2\}}(q_n-q_{n-1})}{\sum_{q_{n-1}=1}^{T}\sum_{q_n=q_{n-1}}^{T}\frac{1}{\sqrt{2\pi}\sigma}\,\mathrm{e}^{-\{[(q_n-q_{n-1})-\mu]^2/2\sigma^2\}}}, \tag{C.1}$$

$$\sigma^2(l) = \frac{\sum_{q_{n-1}=1}^{T}\sum_{q_n=q_{n-1}}^{T}\frac{1}{\sqrt{2\pi}\sigma}\,\mathrm{e}^{-\{[(q_n-q_{n-1})-\mu]^2/2\sigma^2\}}(q_n-q_{n-1})^2}{\sum_{q_{n-1}=1}^{T}\sum_{q_n=q_{n-1}}^{T}\frac{1}{\sqrt{2\pi}\sigma}\,\mathrm{e}^{-\{[(q_n-q_{n-1})-\mu]^2/2\sigma^2\}}} - \mu^2(l). \tag{C.2}$$

References

[1] C. Bunks, D. Mccarthy, A. Tarik, Condition based maintenance of machines using hidden Markov models, Mechanical Systems and Signal Processing 14 (2000) 597–612.

[2] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (1989) 257–286.

[3] P. Baruah, R.B. Chinnam, HMMs for diagnostics and prognostics in machining processes, International Journal of Production Research 43 (2005) 1275–1293.

[4] A. Ljolie, S.E. Levinson, Development of an acoustic–phonetic hidden Markov model for continuous speech recognition, IEEE Transaction on Signal Processing 39 (1991) 29–39.

[5] M. Ostendorf, Stochastic segment model for phoneme-based continuous speech recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing 37 (1989) 1857–1869.

[6] A. Kannan, M. Ostendorf, Comparison of trajectory and mixture modeling in segment-based word recognition, in: Proceedings of the 1993 IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, Minneapolis, MN, 1993, pp. 327–330.

[7] M.Y. Chen, A. Kundu, J. Zhou, Off-line handwritten work recognition using a hidden Markov model type stochastic network, IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (1994) 481–496.

[8] M.Y. Chen, A. Kundu, S.N. Srihari, Variable duration hidden Markov model and morphological segmentation for handwritten word recognition, IEEE Transactions on Image Processing 4 (1995) 1675–1688.

[9] L. Wang, M.G. Mehrabi, E. Kannatey-Asibu, Hidden Markov model-based tool wear monitoring in machining, ASME Journal of Manufacturing Science and Engineering 124 (2002) 651–658.

[10] L. Atlas, M. Ostendorf, G.D. Bernard, Hidden Markov models for monitoring machining tool-wear, in: Proceedings of the 1993 IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 6, Istanbul, Turkey, 2000, pp. 3887–3890.

[11] H.M. Ertunc, K.A. Loparo, A decision fusion algorithm for tool wear condition monitoring in drilling, International Journal of Machine Tools and Manufacture 41 (2001) 1347–1362.

[12] H.M. Ertunc, K.A. Loparo, H. Ocak, Tool wear condition monitoring in drilling operations using hidden Markov models (HMMs), International Journal of Machine Tools and Manufacture 41 (2001) 1363–1384.

[13] C.D. Begg, T. Merdes, C. Byington, K. Maynard, Dynamic modeling for mechanical diagnostics and prognostics, in: Proceedings of the 1999 Maintenance and Reliability Conference, Tennessee, 1999, pp. 22.01–22.13.

[14] M.J. Roemer, G.J. Kacprzynski, Advanced diagnostics and prognostics for gas turbine engine risk assessment, in: Proceedings of the 2000 IEEE International Conference on Aerospace Conference, vol. 6, Big Sky, Montana, 2000, pp. 345–354.

[15] M.J. Roemer, E.O. Nwadiogbu, G. Bloor, Development of diagnostic and prognostic technologies for aerospace health management applications, in: Proceedings of the 2001 IEEE International Conference on Aerospace Conference, vol. 6, Big Sky, Montana, 2001, pp. 63139–63147.

[16] C. Kwan, X. Zhang, R. Xu, L. Haynes, A novel approach to fault diagnostics and prognostics, in: Proceedings of the 2003 IEEE International Conference on Robotics and Automation, vol. 1, Taipei, Taiwan, 2003, pp. 604–609.

[17] P.D. McFadden, J.D. Smith, Vibration monitoring of rolling element bearings by high frequency resonance technique—a review, Tribology International 77 (1984) 3–10.

[18] M.J. Russell, R.K. Moore, Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition, in: Proceedings of the 1985 IEEE International Conference on Acoustics, Speech and Signal Processing, Tampa, FL, 1985, pp. 5–8.

[19] L.E. Baum, J.A. Egon, An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology, Bulletin of the American Mathematical Society 73 (1967) 360–363.

[20] A.J. Viterbi, Error bounds for convolutional codes and an asymptotically optimal decoding algorithm, IEEE Transaction on Information Theory 13 (1967) 260–269.

[21] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society 39 (1977) 1–38.

[22] J.D. Ferguson, Variable duration models for speech, in: Proceedings of the 1980 Symposium on the Application of Hidden Markov Models to Text and Speech, Princeton, NJ, 1980, pp. 143–179.

[23] D. He, D. Wang, A. Babayan, Q. Zhang, Intelligent equipment health diagnosis and prognosis using wavelet, in: Proceedings of the 2002 Conference on Automation Technology for Off-road Equipment, Chicago, IL, 2002, pp. 77–88.

[24] P. Pawelski, D. He, Vibration based pump health monitoring, Journal of Commercial Vehicles 2 (2005) 636–639.

Shaft diagnostics and prognostics development and evaluation using damaged dynamic simulation

S Wu and **D He***
Department of Mechanical and Industrial Engineering, University of Illinois at Chicago, Chicago, Illinois, USA.

The manuscript was received on 24 September 2007 and was accepted after revision for publication on 5 February 2008.

DOI: 10.1243/1748006XJRR108

Abstract: This paper presents a shaft damage dynamic simulation methodology for development and evaluation of effective diagnostic and prognostic algorithms. The methodology uses dynamic shaft vibration models to simulate vibration signals with arbitrary shaft and operating conditions. These simulated vibration signals are used to develop and validate different diagnostic and prognostic algorithms. A simulation study to demonstrate the application of the methodology is provided. In comparison with traditional methods that acquire data from test rig or operational machines, the presented methodology has the following three major benefits. (1) It is able to simulate signals under any shaft and operating conditions (crack size, speed, torque), which is important because of the cost and time required to gather comprehensive operational and test bench data for algorithm development and validation. (2) The methodology could capture nuance for tiny faults without the contamination of extraneous variables, such as ambient noise, test stand dynamics, etc., that are not present in the actual case. Accuracy in this regard is crucial when using vibration data to diagnose faults, for instance to distinguish the effect of noise from a minor crack. (3) It may approach the true signatures of naturally occurring and growing faults more closely by comparison with artificially seeded fault tests or highly accelerated fatigue tests.

Keywords: diagnostics, prognostics, shaft damage dynamic simulation

1 INTRODUCTION

A drivetrain transmission system is one of the most fundamental and important parts of rotorcraft. Shafts are critical components of drivetrain transmission systems, to transfer torque or rotational force from the motor to turn the rotors in a helicopter. Astridge [**1**] reviewed documentation on helicopter accidents in the worldwide civil fleet and showed that 22 per cent of all airworthiness-related accidents (causing death or serious injury, or resulting in loss or substantial damage to the aircraft) since 1956 were attributed to transmission system failures. The other major causes were engines (28 per cent) and rotors (27 per cent). A further breakdown of transmission-related accidents by component was provided to show that approximately 55 per cent of all transmission-related helicopter accidents were caused by shafts.

The ability to monitor the health of shafts accurately can significantly enhance the predictive maintenance task of a drivetrain transmission system. Table 1 gives the general analysis of shaft failure modes. Although all the failure modes in Table 1 could occur to shafts in a helicopter, overload and fatigue cracking are two major contributors to all shaft fracture failure in a helicopter [**1**].

Overload occurs when a single load is applied to cause the part to deform or fracture. Each designed shaft has its own endurance limit $\sigma_{\text{endurance}}$. Given the load σ on the shaft, overload failures occur as soon as $\sigma > \sigma_{\text{endurance}}$. In addition to this, violation of the operation manual by human error could cause overload directly, and incorrect assembly and failure of bearings could lead to overload indirectly by interference also. Fatigue is a progressive localized damage due to fluctuating stresses and strains on the material. Fatigue cracks initiate and propagate in regions where the strain is most severe. The crack

**Corresponding author: Department of Mechanical and Industrial Engineering, University of Illinois at Chicago, 842 West Taylor Street, Chicago, IL 60607, USA. email: davidhe@uic.edu*

Table 1 Summary of shaft failure modes

Failure	Failure mode	Cause	Contributing factors
Shaft fracture	Overload	Interference	Incorrect assembly Bearing failure
		Operational	Operator's mistake
	Wear, abrasion, and erosion	Loss of material by mechanical removal	Surface contact Surface asperities Embedding of debris in surface Melting and oxidation at the interface of the mating surfaces owing to frictional heating
	Corrosion	Loss of material by, for example, pitting, fretting	Electrical or biological action of the corrosion
	Corrosion-influenced fatigue	Material fatigue strength decrease by loss of material	
	Fatigue cracking	Imbalance	Displacement Misalignment Bent shaft
		Crack	Dynamic loads Creep
	Brittle fracture/ductile fracture	Stress rupture Hydrogen embrittlement	Sudden severe overload Chemical action

takes measurable time to progress across the fracture face. Typically, the process of fatigue consists of three stages.

1. The fatigue leads to an initial crack on the surface of the part.
2. The crack or cracks propagate until the remaining shaft cross-section is too weak to carry the load.
3. Final sudden fracture of the remaining cross-section occurs. However, for brittle or ductile fracture, it goes directly to the third stage, more dangerous than fatigue, and hence it is more difficult to monitor health.

Shaft fatigue failure starts with a crack. When a crack propagates over time to the critical size, the shaft finally ruptures, which normally leads to a catastrophic event. Two types of crack are found thus far: surface cracks and in-depth cracks. Surface cracks are the cracks formed on the surface of the shaft. In-depth cracks are the cracks in the shaft. The first type of crack is more common than the second type, and also relatively easy to observe and detect. The cracks discussed in this paper are generally surface cracks. Based on their geometries, surface cracks can be classified as transverse cracks, longitudinal cracks, or slant cracks. Transverse cracks are cracks perpendicular to the shaft axis; they are the most common and most serious, as they reduce the cross-section and thereby weaken the rotor. Longitudinal cracks are cracks parallel to the shaft axis, and slant cracks are cracks in between these two types of crack. Most research focuses on the detection of such transverse cracks. A typical failure face of a shaft is shown in Fig. 1, where the three progression stages are clearly illustrated. The crack initiated at the failure origin, slowly grew, and fractured in the instantaneous zone. Interpretation of this failure face can also disclose the forces that caused the crack, the amount of time elapsed from initiation to final failure, the relative size and type of the load, and the severity of the stress concentrations. The rate at which the crack grows across the face of the part varies with the load on the part. It may take only a few cycles, but in most industrial applications it takes millions of stress applications before the part finally breaks.

There are two main mechanical causes of fatigue failure. The first is imbalance, and the second is cracking.

Failures may take place owing to imbalance. Imbalance occurs when the rotational axis of a shaft and the mass centre of the shaft assembly do not coincide. Shaft imbalance may originate from shaft displacement or misalignment, or its own bent shape. Displacement, misalignment, and/or bent shafts all come from one of the following errors: installation error, manufacturing error, or maintenance error. These errors are not in themselves normally considered to be failures. However, if they continue uncorrected, they can cause distress to other rotating components and fatigue cracking of the shaft itself owing to excessive, non-uniform dynamic loads. All of them can cause vibration, resulting in fatigue failure cracking of the shaft. From this point of view, the principal failure mode that occurs in shafts is cracking. However, the factors contributing to

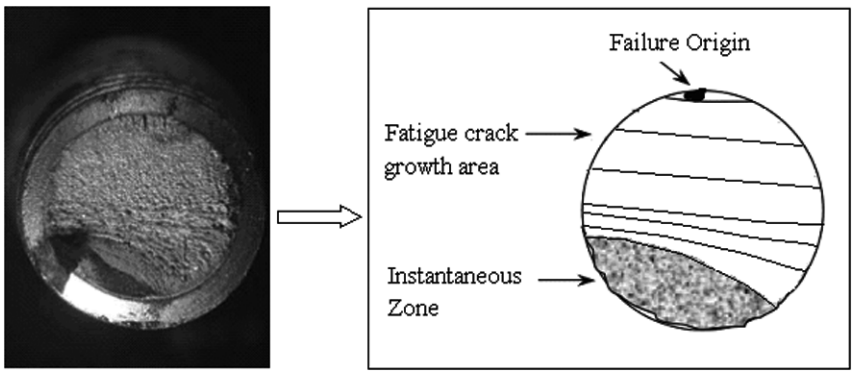


Fig. 1 Cross-section of a typical shaft fatigue failure surface

cracking are manifold. Dynamic loads are the fundamental factor no matter in what kind of environment the shaft is working and what other causes of cracking it has, because cracking never occurs without cyclic load. The shaft fatigue failures also can be classified as bending fatigue, torsional fatigue, and axial fatigue. In the case of axial fatigue, the bearing carrying the load will undergo fatigue (contact fatigue) before the shaft does. This is usually evidenced by spalling of the bearing raceways. In the bending mode, almost all failures are considered 'rotational', with the stress fluctuating or alternating between tension and compression. A thrust load is imposed on the shaft additionally, which can accelerate the speed of fatigue failure beyond the main influence of torsion load. Torsional fatigue is associated with the amount of shaft torque present and the transmitted load. Torque can informally be thought of as 'rotational force'. For the case of the rotating shafts in a helicopter, with an environment of high speeds and high torques, torsional fatigue becomes the main mode, i.e. it is the main mode determining the useful life of the shaft.

Diagnostics and prognostics are two major capabilities for detecting and predicting the health status of a component or system, such as a shaft. Next-generation diagnostic and prognostic systems demand a more accurate and robust diagnostics and prognostics capability to adapt to various operating conditions. Thus, to meet the needs of applicable diagnostic and prognostic systems, developed methods must be tested and evaluated under all kinds of operating conditions to ensure their effectiveness and applicability. There are two alternative ways to obtain data to evaluate the developed approaches. The first one is to use traditional methods that acquire data from test rig or operating machines. The second way is through simulation which is based on physics models. To date, validation tests of diagnostic and prognostic approaches have mainly been conducted by using the empirical data acquired from traditional methods, which has been a bottleneck in developing advanced diagnostic and prognostic capability. For example, many effective model-based diagnostic and prognostic algorithms need various fault data to train the models, including neural networks (NNs) [**2**], hidden Markov models (HMMs) [**3**, **4**], and related approaches such as advanced hidden Markov models (AHMMs) [**5**] and hidden semi-Markov models (HSMMs) [**6**]. However, limited data, especially fault data, have restricted the full evaluation of these developed models. In developing other types of diagnostic and prognostic algorithm, such as data-driven algorithms [**7**, **8**, **9**], statistical reliability based algorithms [**10**, **11**], state estimator based prognostic algorithms [**12**, **13**], and so on, the absence of rich empirical data has an even more negative impact on the performance of the algorithms. The purpose of this paper is to present a damage dynamic simulation method that can be used as a test and evaluation platform for diagnostic and prognostic algorithms. Using damage dynamic models, it is possible to simulate the damage dynamic behaviour of a shaft with arbitrary conditions that are difficult to obtain or control. Additionally, there are various reasons for carrying out simulation instead of fatigue tests.

1. The cost and time to gain all needed empirical data is prohibitively expensive. To study the behaviour difference between normal condition and fault condition, various empirical data have to be analysed. Traditionally, empirical data used to determine features and develop detection and prediction algorithms are obtained from test stands or operational machines under varying operational and fault conditions. These tests are expensive and time consuming. Therefore, the volume and type of empirical data are strictly constricted by these two real-world factors.
2. Full coverage of all fault scenarios is impractical. In real operations, faults occur rarely, and it is impossible to observe all possible faults in limited operation tests. Fault conditions, such as crack location and size, and operating and environmental parameters, such as imbalance,

misalignment, shaft elasticity, gear eccentricity, temperature, load history, etc., are difficult to control practically in test stands.

3. There are some factors that will generate inconsistency between experimental data and actual data. For example, test stands allow either artificially seeded fault tests or highly accelerated fatigue tests. They may not represent the true signatures of naturally occurring faults or the true response of the system when mounted on the actual vehicle. Some variables that influence the collected data, such as ambient noise, test stand dynamics, etc., are usually at variance with actual scenarios. Hence, inconsistencies between test data and actual machine data exist in practice.

The remainder of this paper is organized as follows. Section 2 describes the dynamic simulation methodology. Section 3 provides a simulation case study for shaft diagnostics and prognostics development and evaluation to demonstrate the applicability of the simulation methodology. Finally, section 4 concludes the paper.

2 DAMAGE DYNAMICS BASED SIMULATION METHODOLOGY

2.1 Framework of the methodology

The framework of the proposed methodology is shown in Fig. 2. It consists of two major parts. In the first part, damage dynamic models are used to generate simulated vibration response given the shaft and operating conditions, and the simulated vibration is used to develop and evaluate the diagnostic algorithms. The diagnostic algorithms will determine the condition of the shaft on the basis of the vibration signals: cracked or uncracked. The condition of the shaft will be input to the second part. In the second part, shaft life models, including a prediction shaft life model based on Miner's law and the non-linear Walker equation, are used to simulate the life progression of the shaft, and the simulated progression is used to develop and evaluate the prognostic algorithms. If initialization of shaft cracking is detected, the shaft life model is applied to calculate the remaining useful life (RUL) of the cracked shaft from crack size a_0 to defined failure size a_f, denoted as RUL_2. Otherwise, Miner's law is used to calculate the RUL of the uncracked shaft, denoted as RUL_1, and the life model is used to calculate RUL_2. Therefore, the total RUL of an uncracked shaft is the sum of the RULs of each stage, i.e. $RUL = RUL_1 + RUL_2$. These two parts are integrated physically by crack characteristics, such as crack size. Consequently, measured parameters, such as vibrations, torque, rotational speed, and time horizon, are fully taken into consideration simultaneously for effective diagnostics and prognostics.

2.2 Shaft damage dynamic vibration model

The dynamic behaviour of cracked rotors has been researched since the early 1970s. A comprehensive review is provided in reference [**14**]. Wauer [**15**], Gasch [**16**], and Edwards [**17**] presented three excellent reviews in the field of dynamics of cracked rotors and different procedures to diagnose fracture damage. The dynamic system adopted is a simple hinge model that comprises a rotating shaft with a transverse crack

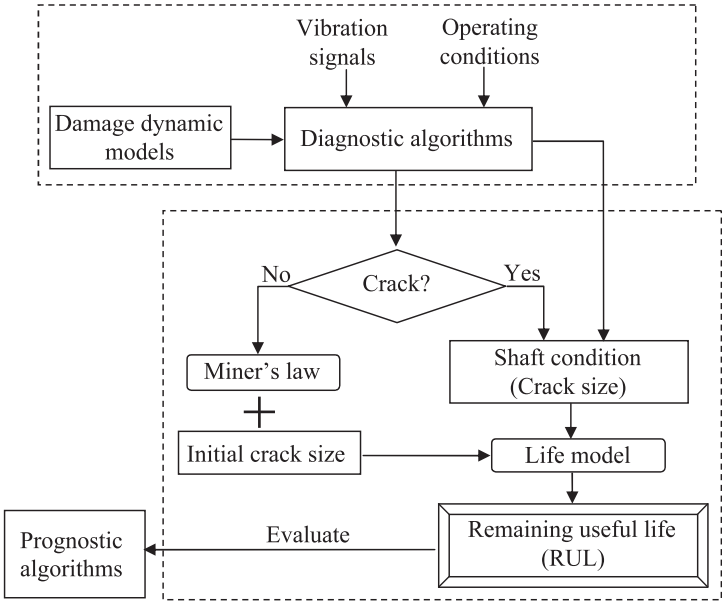


Fig. 2 Framework of the damage physics based dynamic simulation methodology

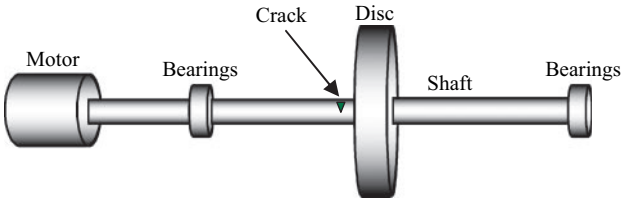


Fig. 3 Schematic of the cracked shaft dynamic system

and a rotor disc. This model only has two degrees of freedom: rotational and vertical bending. A schematic of the modelled dynamic system is shown in Fig. 3. For a crack size less than or equal to half the diameter of the shaft, it has been shown that the simple hinge model is good enough to represent the cyclic stiffness variables [**16**]. If the crack size is larger than half the diameter of the shaft, the crack model of Mayes and Davies [**18**, **19**] is better than the hinge model.

Theoretically, the sum of the periodic forced excitation due to the imbalance and the excitation due to cracking can be derived from the general equation of motion and expressed as a function of eccentricity and change in stiffness, as shown in reference [**16**].

2.3 Shaft life estimation

As mentioned before, the fatigue life of a shaft can be decomposed into two phases: prior-to-crack initialization and after-crack initialization. In the phase of prior-to-crack initialization, Miner's law can be used to calculate the proportion of fatigue life under variable loading, where the failure life is defined by the initialization of a crack. Miner's law is used for calculating the 'safe life' of a component and assumes that fatigue damage accumulates in a linear manner. In a helicopter transmission system, the load on the shaft, σ, is the torque τ. Complex torque history could be decomposed into k series of torque magnitudes by the rainflow technique [**20**]; torque keeps constant in each series, the cumulative torque cycles of each torque level τ_i are denoted as n_i, $1 \leqslant i \leqslant k$, and then the damage level of the shaft, φ, can be written as

$$\varphi = \sum_{i=1}^{k} \frac{n_i}{N_i} \tag{1}$$

where N_i is the number of cycles to failure of a constant torque τ_i and can be determined by the S–N curve. Then, the remaining number of cycles permissible at τ_i is

$$N_{T_i} = N_i(1-\varphi) \tag{2}$$

Miner's law has been widely accepted and used in industry and military applications, and more details can be found in references [**21**] to [**23**].

When $\varphi = 1$, a crack initializes, and Miner's law is no longer useful, as the crack grows non-linearly. It has been found that fatigue crack propagation consists of three distinct stages of crack growth, as shown in Fig. 1, and in each stage a unique crack propagation equation should be used [**24**]. Since Paris [**25**, **26**], most research on fatigue crack growth has related the crack growth properties to the stress intensity factor. The crack growth rate $\mathrm{d}a/\mathrm{d}N$ can always be determined by using the stress intensity factor amplitude ΔK, and it can be written as [**25**]

$$\frac{\mathrm{d}a}{\mathrm{d}N} = C(\Delta K)^{\alpha} \tag{3}$$

where C and α are parameters determined by the material, circumstance, loads, and so on. For example, α is normally approximately equal to 3 for steel. The amplitude of the stress intensity factor ΔK could be obtained from the equation

$$\Delta K = Y\,\Delta\sigma\sqrt{\pi a} \tag{4}$$

where Y is a geometry factor depending on the loading and cracked-body configuration, a is crack size, and $\Delta\sigma$ is the load stress amplitude, which is equal to the difference between maximum stress and minimum stress, i.e. $\Delta\sigma = \sigma_{\max} - \sigma_{\min}$. Paris' law and the formulae derived on the basis of this law provide practical solutions for many engineering problems. The law can successfully describe the experimental data for long cracks under small-scale yielding and constant-amplitude loading. Modifications are needed to consider the R-ratio effect [**27**], threshold limits [**28**, **29**], variable-amplitude loading [**30**], the overload effect [**31**], and small cracks [**32**]. For example, without considering the influence of stress ratio R on $\mathrm{d}a/\mathrm{d}N$ some limitations will be imposed on the usage of the Paris equation, as Paris' law does not account for crack growth rate at both low and high levels of ΔK. In fact, experiments indicated that, at high ΔK values, the crack growth rate $\mathrm{d}a/\mathrm{d}N$ increases with increasing stress ratio R and has little sensitivity at lower ΔK levels [**33**]. Thus, Walker's equation [**34**] is chosen as the crack propagation model instead of Paris' law, taking the factor stress ratio R into consideration. The stress ratio R is defined as the ratio between minimum stress and maximum stress during the rotation process, $R = \sigma_{\min} / \sigma_{\max}$.

When the stress ratio R is considered, the relationship between $\mathrm{d}a/\mathrm{d}N$ and ΔK can often be expressed by Walker's equation [**34**]

$$\frac{\mathrm{d}a}{\mathrm{d}N} = C\left[\frac{\Delta K}{(1-R)^{\gamma}}\right]^{\alpha} \tag{5}$$

where C, α, and γ are determined by the material, circumstance, frequency, temperature, and so on.

Hence, the remaining cycles N_T, starting from the initial crack size a_0 to the critical failure crack size

a_f, can be computed as

$$N_T = \int_{a_0}^{a_f} \frac{da}{C\left\{\Delta K/[(1-R)^\gamma]\right\}^\alpha} = \int_{a_0}^{a_f} \frac{da}{C\left\{Y\Delta\sigma\sqrt{\pi a}/[(1-R)^\gamma]\right\}^\alpha} = \frac{C^{-1}}{\alpha/2-1}\left(\frac{Y\,\Delta\sigma\sqrt{\pi}}{(1-R)^\gamma}\right)^{-\alpha}(a_0^{1-\alpha/2}-a_f^{1-\alpha/2}) \quad (6)$$

provided that $\alpha > 2$ and C, Y, and γ are constant (see reference [**35**]). If the initial crack size is much less than the final size, $\alpha_0 \ll \alpha_f$, then equation (6) could be expressed approximately as

$$N_T \cong \frac{C^{-1}}{\alpha/2-1}\left(\frac{Y\,\Delta\sigma\sqrt{\pi}}{(1-R)^\gamma}\right)^{-\alpha} a_0^{1-\alpha/2} \quad (7)$$

Note that in equation (7) the load stress amplitude is the torque load because a helicopter rotor shaft mainly suffers mode III (out-of-plane shear, tearing), which is caused by torque, as opposed to mode I (tension, opening) and mode II (in-plane shear, sliding) [**36**].

Equation (7) can be used to predict the fatigue life, given the crack size. However, in practice, crack size is usually unknown and not able to be measured directly. Therefore, establishing N_T or the RUL with a parameter that can be directly monitored is critically important and meaningful.

Next, a comprehensive model that directly links N_T with vibration response is established. The model is then used as the main model in the simulation.

2.4 Comprehensive model

The two key parameters that can be used to establish the comprehensive model are stiffness and crack size. Factors influencing stiffness include material modulus, the structural configuration of load-transmitting components (bars, plates, rods, shells), the mode of loading, and the contact deformation between mating parts (boundary condition). Cracking affects the shaft structural configuration. As a crack propagates across the shaft, the remaining cross-section becomes smaller, and the bending stiffness of the shaft decreases. In references [**37**] and [**38**], an expression that takes into account the effects of crack size *a,* shear deformation, and rotary inertia for calculating the stiffness s has been developed and proved. The expression for stiffness can be written as

$$s = \frac{48\ (I/L^3)E}{1 + 12q^2 + (6d/L)V(a/\Phi)} \quad (8)$$

where L is the span of the shaft, E is Young's modulus of elasticity, the moment of inertia of the intact rotating shaft is $I = (\pi/64)\Phi^4$, $q^2 = (EI)/(kG'AL^2) = \{[2(1+\nu)]/k\}r^2$, $r^2 = I/(AL^2)$, and the shear modulus is $G' = E/[2(1+\nu)]$. When the cross-section is circular, $k = [6(1+\nu)^2]/(7+12\nu+4\nu^2)$, where ν is Poisson's ratio. $V(a/\Phi)$ is given by Bakker [**39**].

Note that, from equation (8), when the crack size $a = 0$, the bending stiffness of the shaft with valid length L, ignoring the shear deformation, can be computed as

$$s_0 = 48\frac{I}{L^3}E \quad (9)$$

Given a crack size a_0, from equations (8) and (9), the change in stiffness Δs_ξ can be computed as

$$\Delta s_\xi = s_0 - s = 48\frac{I}{L^3}E - \frac{48\ (I/L^3)E}{1 + 12q^2 + [(6d)/L]\,V(a_0/\Phi)} = 48\frac{I}{L^3}E\left(\frac{12q^2 + [(6d)/L]V(a_0/\Phi)}{1 + 12q^2 + [(6d)/L]V(a_0/\Phi)}\right)$$

Hence, based on reference [**16**], the overall excitations due to imbalance and cracking can be expressed as

$$\Delta r(t) = \varepsilon\left(\frac{\eta^2}{1-\eta^2+2\mathrm{j}D\eta}\right)\mathrm{e}^{\mathrm{j}(\Omega t+\beta)} + \lambda\sum_{k=-3}^{k=+3}\frac{b_k\mathrm{e}^{\mathrm{j}k\Omega t}}{1-k^2\eta^2+2\mathrm{j}Dk\eta} \quad (10)$$

where constant item

$$\lambda = \frac{mgL^3}{48EI}\left(\frac{12q^2 + [(6d)/L]V(a_0/\Phi)}{1 + 12q^2 + [(6d)/L]V(a_0/\Phi)}\right)$$

given crack size a_0.

Equation (10) shows that the forced excitation is decided by many factors, such as the elasticity modulus E, the shear modulus G', damping d, geometry parameters, mass, crack size, rotational speed and so on, which also means that the shaft dynamic model already takes these factors into consideration. Next, a simulation case study to demonstrate the application of the methodology is described.

3 SIMULATION CASE STUDY

3.1 System description

The system in the study is based on the schematic shown in Fig. 3, and a transverse fatigue crack was created artificially at the midspan of the shaft. Different crack sizes were chosen for model validation and simulation. The parameters of a cracked rotating shaft are given in Table 2.

3.2 Model validation

A validation experiment test was conducted to validate the theoretical physics vibration model of a shaft crack. In the test, vibration data were acquired from an experimental set-up designed to perform vibration testing of shafts with a crack. The equipment used in this experiment was a bearing balance simulator (see Fig. 4). The simulator is specifically

Table 2 Parameters of shaft

Parameters	Value	Unit
Length L	0.240 35	m
Diameter ϕ	0.015 82	m
Mass of shaft m_1	0.566 74	kg
Mass of disc m_2	6.2868	kg
Coefficient of damping d	0.001	—
Poisson's ratio ν	0.29	—
Young's modulus of elasticity E	200×10^9	kg/(m s^2)

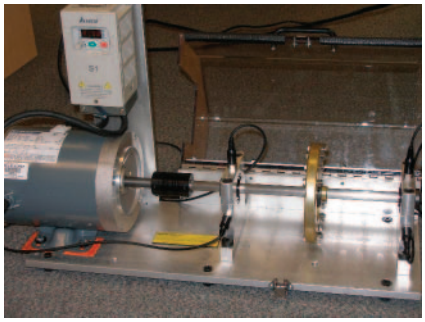


Fig. 4 Bearing balance simulator

designed to demonstrate and support the study of bearing or shaft faults under controlled conditions. It is a variable-speed machine that can be used to generate each type of fault individually or in combination, providing a stable platform for study. Four sensors are distributed in two supports. Each support has a horizontal sensor and a vertical sensor. The left support is connected to channels 1 and 4, and the right support is connected to channels 2 and 3. Thus, the vibration of each support is the combination of these two components into a complex vector.

Five identical shafts were prepared to perform the test. Among them, one was intact, and the others had a crack with a depth of 2, 3, 5, and 7 mm respectively. During the test, each shaft was installed to run in sequence. The shaft was tested at a constant rotational speed of 4200 r/min. A PC-based data acquisition system was used continuously to acquire and store shaft vibration spectra with a sampling rate of 5120 Hz. The record time per block was 6.4 s. The FFT frequency limit was 2 kHz. Acquired raw data were represented by displacement. It has been shown that, for low rotational speed, when the unbalance effects are negligible, the vibration at $1\times\Omega$ is due to the crack [**40**]. Therefore, the first shaft order was used to validate the vibration model due to cracking.

The results from experiments and theoretical values of the test are provided in Table 3. It is clear that the first shaft order grows as crack size increases.

As can be seen from Table 3, the first shaft order computed from the physical model is consistently proportional to that obtained from the fault simulation experiment. This proportion can be obtained by computing the ratio of the first shaft order

Table 3 Model validation results

Crack size (mm)	SO_1 on left support (mm)	Excited SO_1 due to crack experimentally (mm)	Excited SO_1 due to crack theoretically (mm)	Ratio*
0	9.24×10^{-8}	0	0	—
2	1.47×10^{-4}	1.47×10^{-4}	3.02×10^{-4}	0.49
3	1.79×10^{-4}	1.79×10^{-4}	6.29×10^{-4}	0.28
5	4.39×10^{-4}	4.39×10^{-4}	1.59×10^{-3}	0.28
7	8.52×10^{-4}	8.52×10^{-4}	2.99×10^{-3}	0.29

*Ratio = (excited SO_1 due to crack experimentally)/(excited SO_1 due to crack theoretically).

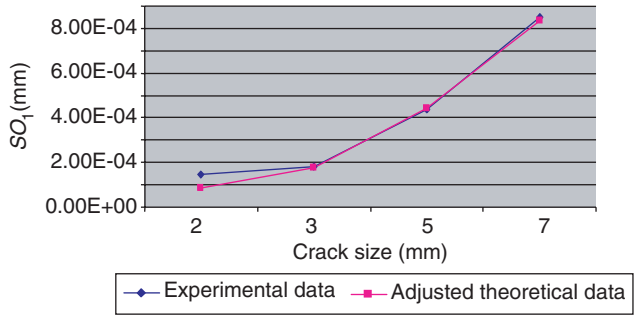


Fig. 5 Experimental first shaft order versus adjusted theoretical data due to cracking

obtained from the fault simulation experiment over the first shaft order computed from the physical model. Again, the results are shown in Table 3. If the theoretical value is adjusted by a factor of 0.28, the adjusted theoretical data fit the experimental data very closely, as shown in Fig. 5. The reason for adjusting the theoretical value is because the theoretical value is the vibration at the midspan of the shaft, while the measured vibration is at the left support of the rotor. Therefore, the experimental values are supposed to be less than the theoretical values.

The difference between the experimental values and the theoretical value is multifold. Since five shafts were used to validate the vibration model due to cracking, minor variance between each shaft and variance of assembly conditions did exist. Both of these would affect the vibration. Meanwhile, when the crack size is small, the induced vibration is relatively small. The measured vibration due to small crack size is affected by noise more than that due to large crack size.

3.3 Evaluation of diagnostics

In this case study, the shaft parameters used for simulation are the same as those in Table 2, and seven diagnostic features extracted from vibration were tested. These features were: shaft orders 1, 2, and 3, kurtosis (fourth moment of signal), M6 (sixth moment of signal), root-mean-square (RMS), and peak-to-peak value.

In the study, crack sizes ranging from 0 to $a_f = \Phi/2$ with a fixed increment were simulated for each specific loading condition, where Φ is the diameter of the shaft. Note that crack size is defined such that 0 represents a normal shaft where there is no crack, while crack size $\Phi/2$ represents a crack that runs half the diameter of the shaft. The RUL is defined as the useful life from the current moment to the time when crack size grows to half the diameter of the shaft. Vibration signals were generated under the influence of various simulated crack sizes and operation conditions. Based on the definitions of features and the dynamic vibration model, comparison simulation tests under different operating conditions were conducted on these features.

The first condition to simulate was setting different shaft rotational speeds given a constant input torque of 10 Nm. Four different rotational speeds – 600, 2400, 4800, and 12 000 r/min – were simulated as the loading conditions for the shaft. Thus, 28 vibration signals were generated (seven crack sizes by four speeds). The second condition to simulate was setting different input torques given a constant speed of 4800 r/min. Three different torques –1, 10, and 100 N m, were simulated as the loading conditions for the shaft. Thus, 21 vibration signals were generated (seven crack sizes by three torques). The detailed simulation results can be found in reference [**41**].

3.3.1 *Diagnostic performance*

Based on the results of the simulation, the following observations can be made.

1. The three shaft orders consistently increase as the crack size increases, regardless of the rotational speed and torque load. The increases in SO_1 due to increases in crack size at low speed are more significant than the increases in SO_2 and SO_3. However, the increases in SO_1 due to increases in crack size at high speed are smaller than the increases in SO_2 and SO_3. Rotating speed impacts more on the sensitivity of SO_2 and SO_3 to increase in crack size than on that of SO_1. Given a constant speed, it can be observed that the torque load has little impact on the increases in shaft orders as the crack size increases.
2. For any given speed or torque load, kurtosis shows little change as the crack size changes.
3. RMS consistently increases as the crack size increases, regardless of the rotational speed and load. However, RMS is more sensitive to crack size at high rotational speed.
4. For any given speed or torque load, M6 shows little change as the crack size changes.
5. Peak-to-peak value consistently increases as the crack size increases, regardless of the rotational speed and torque load.

Based on the above observations, it can be seen that the shaft orders, speed, and peak-to-peak value can be good candidates of diagnostic features for shaft crack detection.

3.3.2 *Robustness of diagnostic features*

In real machinery health-monitoring applications, it is inevitable that the vibration signals are noisy. Hence, in this simulation study, the effects of vibration noise on the performance of the candidate diagnostic features, shaft orders, RMS, and peak-to-peak value were investigated. The condition of the given case was a speed of 600 r/min and a torque of 10 Nm. Robustness was studied by adding various levels of Gaussian noise. The level of the vibration noise was represented by its standard deviation. The robustness can be computed as the percentage deviation of diagnostic feature values at each noise level from the 'true' diagnostic feature values, i.e.

$$\text{percentage deviation} = \frac{\text{noisy feature value} - \text{noise-free feature value}}{\text{noise-free feature value}}$$

The plots of percentage deviation for all selected diagnostic features are provided in Fig. 6.

From Fig. 6 it can be seen that the percentage deviation is higher when the crack size is smaller and the noise level is larger. It can also be seen that the deviation percentage of higher shaft order is greater than that of lower shaft order under the same crack size and noise level. It can therefore be concluded that background noise has more effect on detection of a smaller crack size, which induces smaller vibration, than a larger crack size, and more effect on higher harmonic order. In particular, when background noise swamps the crack signal, shaft order becomes insensitive to the crack. For example, when the noise level is 0.01, all harmonic shaft orders deviate far away from the true value when the noise level is 0. Under the given parameters of the dynamic system and maximum operating conditions, all shaft order values of the signal are below 10^{-2}. It was found that shaft order becomes insensitive to the crack when the noise level reaches up to 10^{-2}.

Compared with shaft orders, noise has much more impact on RMS and peak-to-peak value – it can clearly be seen that both of them have much higher deviation percentage than shaft orders under the same noise level. The rough threshold of background noise level that makes RMS and peak-to-peak value ineffective is around 0.0005 and 0.0001 respectively. However, this value is 0.01 for the first shaft order. That is to say, shaft orders have much better performance than RMS and peak-to-peak value in a noisy environment.

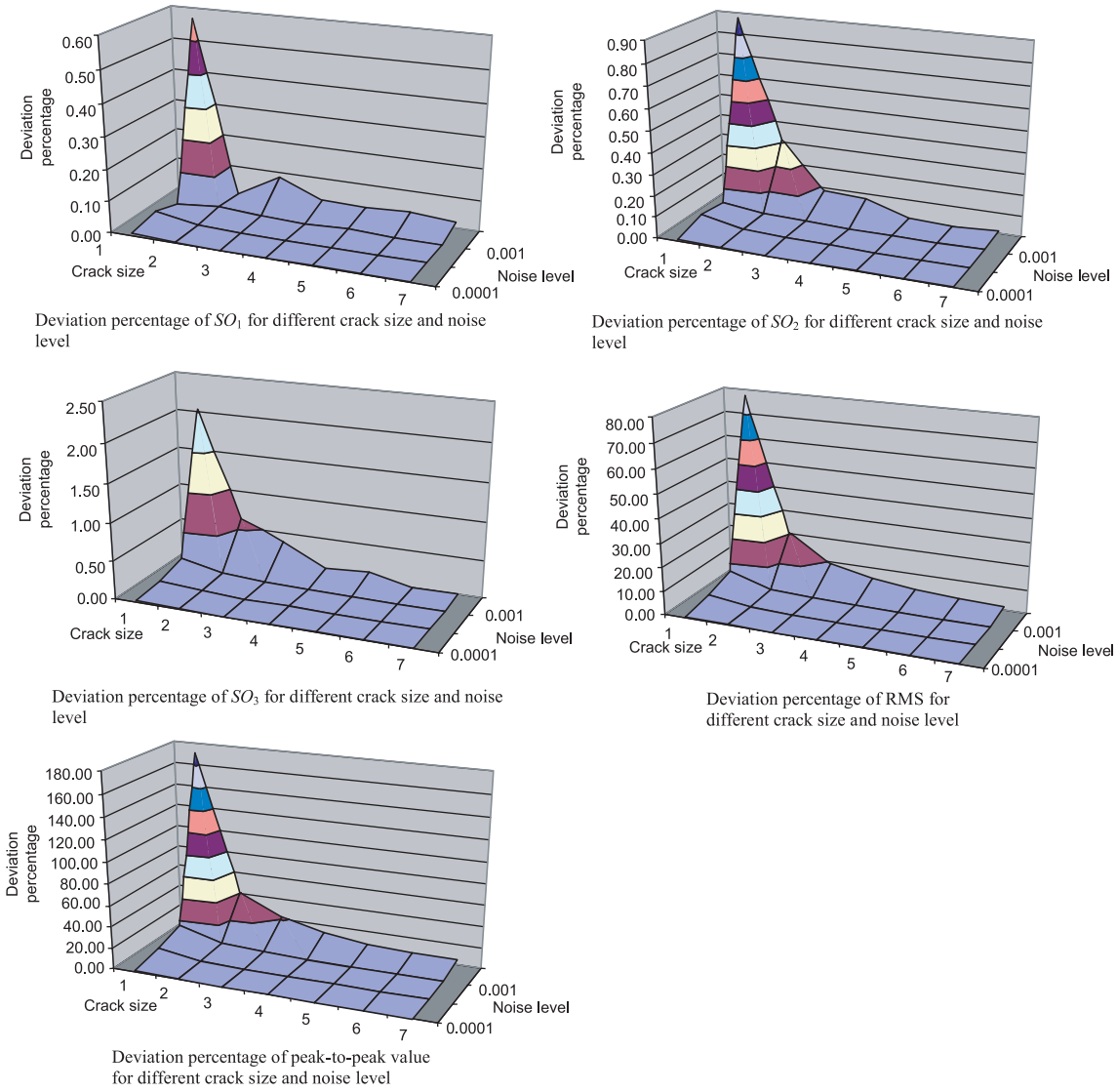


Fig. 6 Deviation percentage of five candidate diagnostic features for different crack sizes and noise levels

Therefore, it can be concluded that shaft orders are the best diagnostic features for shaft order monitoring and detection.

3.4 Development and evaluation of prognostics

In this section, the development and evaluation of prognostics using simulated shaft life progression are explained.

Taking only two failure modes – imbalance and cracking – into consideration, the shaft order of the system vibration SO is the sum of the shaft order of normal condition SO_{norm}, the shaft order due to eccentricity SO_ε and the shaft order due to cracking SO_{crack}

$$SO = SO_{\text{norm}} + SO_\varepsilon + SO_{\text{crack}}$$

Normally, given eccentricity ε initially (in reality, eccentricity always exists no matter how well the shaft is assembled), SO_{norm} and SO_ε do not change over time without human interference. Thus, $SO_{\text{norm}} + SO_\varepsilon$ comprises the base shaft order of the system vibration. When no crack initializes, i.e. $SO_{\text{crack}} = 0$, the shaft order of the system vibration SO is $SO_{\text{norm}} + SO_\varepsilon$. When the crack initializes, SO is increased by SO_{crack}.

Therefore, by monitoring the vibration of a cracked shaft, the first shaft order can be computed by Fourier transform (see Appendix 2 for computation of shaft orders). Knowing SO_1, the constant term λ in equation (10) can be computed (see Appendix 2 for computation of λ using shaft order 1). As the constant λ contains the initial crack size a_0, it is possible to determine a_0 by taking the inverse function of λ as $a_0 = f^{-1}(\lambda)$. Once crack size a_0 is calculated, its associated RUL can be derived on the basis of equation (7).

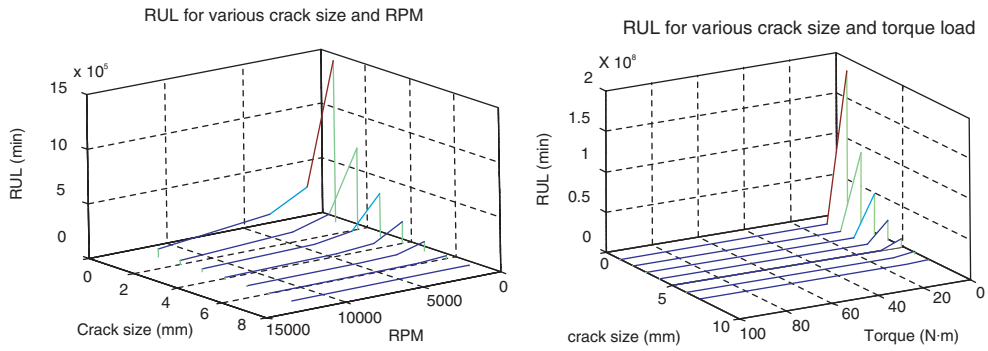


Fig. 7 RUL spectrum fall

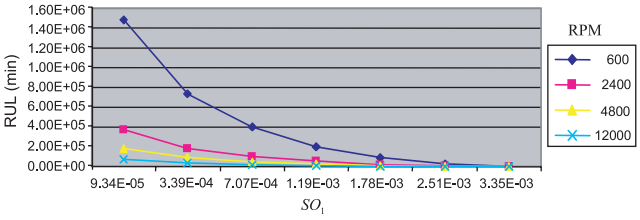


Fig. 8 RUL versus SO_1 at a torque of 10 N m

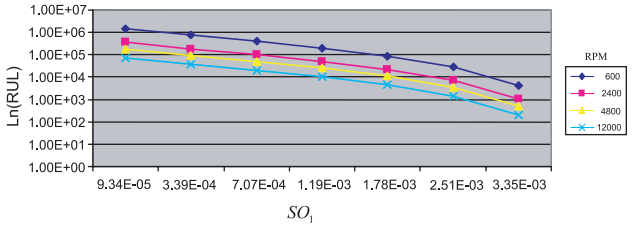


Fig. 9 ln(RUL) versus SO_1 at a torque of 10 N m

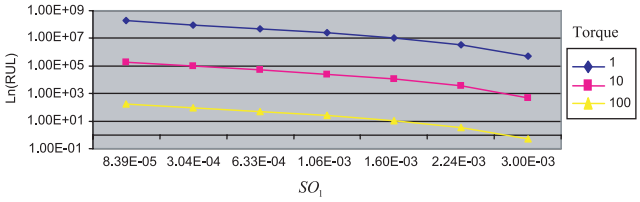


Fig. 10 ln(RUL) versus SO_1 at a speed of 4800 r/min

3.4.1 *Computation of RUL under constant torque and constant speed*

Under the same conditions as in section 3.3.1, the RUL for various crack sizes and rotational speeds and the RUL for various crack sizes and torque loads were computed from life model (7), as shown in Fig. 7. From Fig. 7 it can be seen that the RUL decreases approximately linearly with increase in crack size, but approximately exponentially with increase in speed when the torque is constant. The RUL decreases approximately exponentially with increase in crack size and torque when the rotational speed is constant.

3.4.2 *Relationship between RUL and SO_1*

From the diagnostics evaluation simulation results it can be seen that SO_1 effectively reflects the trend of crack size given changing speed or torque. Additionally, the changes in both cases show regular patterns. Therefore, an attempt was made to establish a simple and effective prognostic model that maps RUL to SO_1. Figure 8 shows a plot of RUL against SO_1 under a torque of 10 N m.

From Fig. 8, it can be seen that the relationship between RUL and SO_1 is approximately exponential. By taking the logarithmic transformation of RUL, a more obvious relationship is shown in Fig. 9.

From Fig. 9 it can be seen that all curves are approximately parallel. These plots imply that the relationship between ln(RUL) and feature SO_1 can be described as a prognostic model $\ln(\mathrm{RUL}) = \lambda f(SO_1)$. Here, λ is a constant depending on the rotational speed. This function can be illustrated by plotting ln(RUL) against SO_1 for a speed of 4800 r/min (Fig. 10).

By generalizing all the findings above, it is found that, in the approximate prognostic model $\ln(\mathrm{RUL}) = \lambda f(\mathrm{SO}_1)$, the factor λ contains information on the influence of torque and rotational speed, SO_1 contains information on crack size, and $f(SO_1)$ is the base prognostic function. Among these, SO_1 is acquired data, but, function $f(SO_1)$ and λ are unknown.

To determine the prognostic model $\ln(\mathrm{RUL}) = \lambda f(SO_1)$, polynomial curve fitting was used. An example is provided to show how it was generated. In this example, the speed was set to 60 r/min and the torque was set to 1 N m. The rule used to evaluate the fitting is the confidence bound. The one that gives the closest confidence bound to the fitted curve has the best fit. Figure 11 demonstrates polynomial curve fittings and their corresponding 95 per cent confidence bounds by changing the degree of the polynomial.

The confidence bound for a fitted coefficient b_i is given by $C = b_i \pm t\sqrt{S_i}$, where t is the inverse of

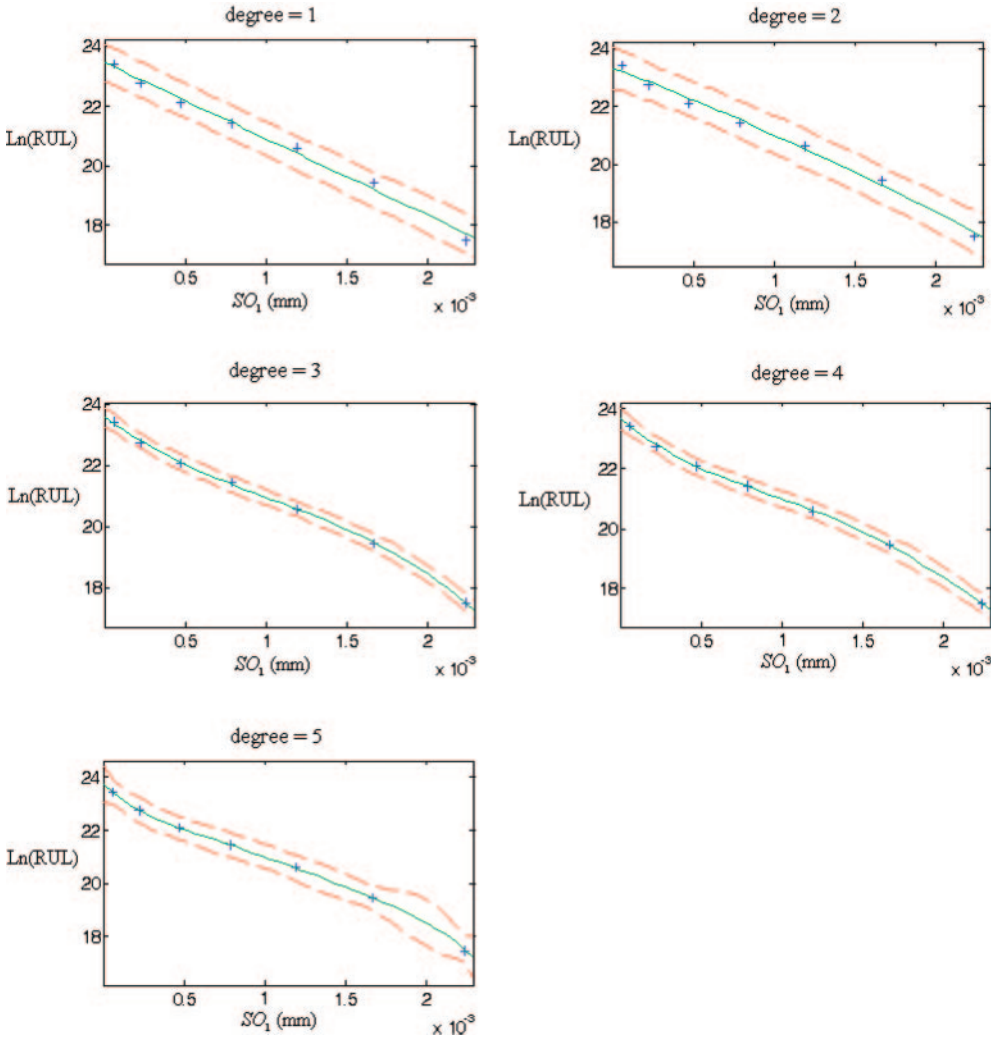


Fig. 11 Polynomial curve fittings and their corresponding confidence bounds

Student's T cumulative distribution function, and S_i is the ith element of $\boldsymbol{S}$, where $\boldsymbol{S}$ is a vector of the diagonal elements from the covariance matrix of the coefficient estimates, $(\mathbf{X}'\mathbf{X})^{-1}s^2$ [**42**]. $\mathbf{X}$ is the design matrix, $\mathbf{X}'$ is the transpose of $\mathbf{X}$, and s^2 is the mean squared error of the estimation. Here, the design matrix $\mathbf{X}$ is composed of SO_1 values, expressed as

$$\begin{bmatrix} x_1^q & x_1^{q-1} & \dots & x_1^0 \\ x_2^q & x_2^{q-1} & \dots & x_2^0 \\ \vdots & \vdots & \vdots & \vdots \\ x_m^q & x_m^{q-1} & \dots & x_m^0 \end{bmatrix}$$

where q is the degree of the fitted polynomial, and m is the total number of SO_1 values.

To find the best-order polynomial curve, degrees of 1–5 were tried (see Fig. 11). A cubic polynomial has the best fit because it has the narrowest confidence bounds. A higher-order polynomial was tried to see if even more precise predictions were possible. For example, a quintic model does have a more precise fit near the data points. However, as measured by the confidence bounds, in the region between the data groups, the uncertainty of prediction rises dramatically. The bulge in the confidence bounds occurs because the data really do not contain enough information to estimate the higher-order polynomial terms precisely, so even interpolation using polynomials can be risky in this case.

After the appropriate order of the polynomial function has been determined, it is easy to derive the coefficients for the polynomial function. The curve fitting of the third-order polynomial function for the given example is

$$\ln(\text{RUL}) = -5.9 \times 10^8 SO_1^3 + 1.9 \times 10^6 SO_1^2 - 3.9 \times 10^3 SO_1 + 24$$

Since the vertical axis is a logarithmic scale of RUL, the final function describing the relationship between RUL and SO_1 for this particular example is expressed as

$$\text{RUL} = e^{\lambda(-5.9 \times 10^8 SO_1^3 + 1.9 \times 10^6 SO_1^2 - 3.9 \times 10^3 SO_1 + 24)} \quad (11)$$

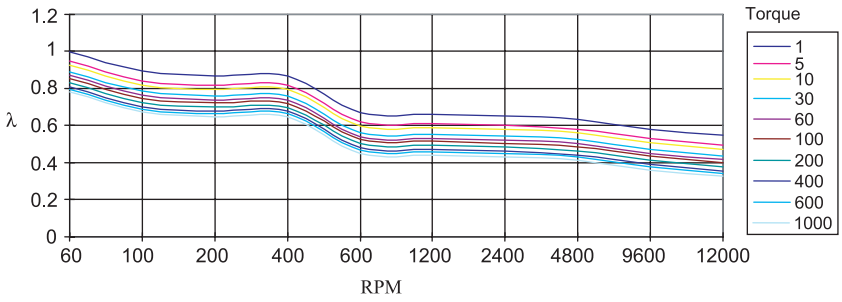


Fig. 12 Coefficient λ versus rotational speed and torque

Note that constant λ in the prognostic model $\ln(\text{RUL}) = \lambda f(SO_1)$ is determined by both rotational speed and torque. For prognostic implementation, the value of λ could be obtained by setting up a base condition in which λ is set equal to 1. In this case, the prognostic model under the base condition is $[\ln(\text{RUL})]_0 = f(SO_1)$. Then the λ value at any given combination of rotational speed and torque can be computed as

$$\lambda = \frac{[\ln(\text{RUL})]}{[\ln(\text{RUL})]_0} = \frac{\lambda f(SO_1)}{f(SO_1)}$$

Therefore, once $f(SO_1)$ is determined for the base condition, the prognostic model, given any other operating condition, can be established by determining λ for that condition as $\lambda f(SO_1)$. Figure 12 shows λ values at different conditions obtained by simulation. As the set base condition in the simulation is a rotational speed of 60 r/min and a torque of 1 N m, $f(SO_1) = -5.9\times10^8 SO_1^3 + 1.9\times10^6 SO_1^2 - 3.9\times10^3 SO_1 + 24$.

3.4.3 *Validation of prognostics*

To test the robustness of the developed prognostic algorithm, a validation test was performed. In this test, vibration signals with different levels of noise were generated, given a shaft crack size of 1 mm and an operation condition of 9600 r/min and 400 N m.

The 'true' RUL can be computed by shaft life model (7) as 1.4489 min. Applying the Fourier transform to the simulated vibration, for example, the SO_1 of the noisy signal with a noise level of 0.001 was computed as 8.88×10^{-5}. Based on the given operating condition, a value of λ approximately equal to 0.4 can be found. Since equation (11) was estimated at the base condition (60 r/min, 1 N m), the RUL of the simulated signal can be computed as

$$\begin{aligned}\text{RUL} &= e^{\lambda f(SO_1)}/\text{rotational speed}\\ &= e^{\lambda(-5.9\times10^8 SO_1^3 + 1.9\times10^6 SO_1^2 - 3.9\times10^3 SO_1 + 24)}/\\ &\qquad\qquad \text{rotational speed}\\ &= e^{0.40\times(-5.9\times10^8\times(8.88\times10^{-5})^3 + 1.9\times10^6\times(8.88\times10^{-5})^2}\\ &\qquad\qquad -3.9\times10^3\times(8.88\times10^{-5})+24)/9600\\ &= 1.3469\ \text{min}\end{aligned}$$

Table 4 Results of the validation test

Noise level	SO_1 (mm)	RUL (min)	Accuracy
0	8.89×10^{-5}	1.3466	0.9294
0.0001	8.96×10^{-5}	1.3453	0.9285
0.0005	8.90×10^{-5}	1.3464	0.9293
0.001	8.88×10^{-5}	1.3469	0.9296
0.01	1.34×10^{-4}	1.2649	0.8730

Therefore, accuracy of prediction is computed as

$$1 - \left|\frac{1.3469 - 1.4489}{1.4489}\right| = 0.9296$$

Similarly, five scenarios with different noise levels were tested. The results of the validation test are given in Table 4.

From Table 4 it can be seen that the accuracy of RUL prediction is around 93 per cent when the noise level is below 0.01. It is also found that a noise level below 0.01 has little impact on the prediction accuracy. However, it suddenly decreases from 93 to 87.30 per cent when the noise level reaches 0.01.

4 CONCLUSIONS

In this paper, a damage physics based dynamic simulation methodology has been presented, and its application to development and evaluation of diagnostic and prognostic algorithms has been demonstrated with a simulation case study.

The simulation methodology is an integration of two main parts: damage dynamic models and shaft life models. Damage dynamic models were used to generate simulated vibration response given shaft and operating conditions, and the simulated vibration was used to develop and evaluate the diagnostic algorithms. Shaft life models were derived on the basis of Miner's law and Walker's equation and used to simulate the life progression of the shaft. The simulated life progression is used to develop and evaluate the prognostic algorithms. A comprehensive model to compute the RUL directly using the first shaft order of the vibration was derived by integrating the damage dynamic models and shaft life models.

A simulation study to demonstrate the application of the methodology to development and evaluation of diagnostic and prognostic algorithms was performed. In this simulation study, laboratory experiments were performed to show that damage dynamic vibration models were able to simulate the actual vibration response of a real shaft system. Several features extracted from vibration response to diagnose shaft cracking under different operational conditions and noise levels were evaluated in the simulation study. The evaluation results show that shaft orders, RMS, and peak-to-peak value increase consistently with increase in crack size, while kurtosis and M6 did not show a consistent response to the change in crack size. The simulation study of effect of noise further reveals that background noise has much less effect on shaft orders than RMS and peak-to-peak value. That is to say, shaft orders had the best performance for shaft crack detection. The study results also show that detection of smaller cracks is more easily affected by background noise than that of larger cracks.

Using the simulated vibration, given shaft and operating conditions and estimated shaft life progression, a simple prognostic model mapping SO_1 to RUL was derived by polynomial curve fitting. The variance between operating conditions can be simplified and represented by a constant factor λ. A validation test has demonstrated that the derived prognostic model is effective, robust, and applicable.

In comparison with traditional methods, which acquire data from test rig or operational machines, the methodology presented has three major benefits.

1. It is able to simulate signals under any shaft and operating conditions (crack size, speed, torque). This is important because of the cost and time required to gather comprehensive operational and test bench data for algorithm development and validation.
2. The methodology can capture nuance for tiny faults without contamination by extraneous variables such as ambient noise, test stand dynamics, etc., that are not present in actual conditions. Accuracy in this regard is crucial when using vibration data to diagnose faults, for instance to distinguish the effect of noise from a minor crack.
3. It can approach the true signatures of naturally occurring and growing faults more closely compared with artificially seeded fault tests or highly accelerated fatigue tests.

ACKNOWLEDGEMENTS

This paper was funded by the Aviation Applied Technology Directorate, Aviation and Missile Research, Development and Engineering Center under Technology Investment Agreement with W911W6-05-2-0003, entitled Rotorcraft Research and Innovation. The authors would like to acknowledge that this research and development was accomplished with the support and guidance of AATD. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Aviation and Missile Research, Development and Engineering Center or the U.S. Government. Special thanks are also given to the anonymous reviewers for their constructive comments on improving the presentation quality of the paper.

REFERENCES

1 **Astridge, D. G.** Helicopter transmissions – design for safety and reliability. *Proc. Instn Mech. Engrs, Part G: J. Aerospace Engineering*, 1989, **203**(G2), 123–138.
2 **Jang, J. S. R., Sun, C. T.,** and **Mizutani, E.** Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence. *MATLAB curriculum series*, 1997, pp. xxvi–614 (Prentice Hall, Upper Saddle River, New Jersey).
3 **Atlas, L., Ostendorf, M.,** and **Bernard, G. D.** Hidden Markov models for monitoring machining tool-wear. In Proceedings of IEEE International Conference on. *Acoustics, speech and signal processing*, Istanbul, Turkey, 5–9 June, 2000, vol. 6, pp. 3887–3890.
4 **Roemer, M. J., Nwadiogbu, E. O.,** and **Bloor, G.** Development of diagnostic and prognostic technologies for aerospace health management applications. In Proceedings of the 2001 IEEE Aerospace Conference, Big Sky, Montana, 2001, .
5 **Dong, M., He, D., Banerjee, P.,** and **Keller, J.** Equipment health diagnosis and prognosis using advanced hidden Markov models. In Proceedings of the 58th Meeting of the Society for Machinery Failure Prevention Technology, Virginia Beach, 25–30 April 2004, pp. 375–384.
6 **Dong, M.** and **He, D.** Hidden semi-Markov models for machinery health diagnosis and prognosis. *Trans. NAMRI/SME*, 2004, **32**, 199–206.
7 **Bechhoefer, E., Bernhard, A., He, D.,** and **Banerjee, P.** Use of hidden semi-Markov models in the prognostics of shaft failure. American Helicopter Society 61st Forum, Phoenix, Arizona, 2005.
8 **Byington, C. S., Watson, M.,** and **Edwards, D.** Data-driven neural network methodology to remaining life predictions for aircraft actuator components. IEEE Aerospace Conference Proceedings, Big Sky, Montana, 2004.
9 **He, D., Wu, S.,** and **Banerjee, P.** Probabilistic model based algorithms for prognostics. In Proceedings of the IEEE Aerospace Conference, Big Sky, Montana, USA, March 2006, pp. 1–10.
10 **Goode, K. B., Moore, J.,** and **Roylance, B. J.** Plant machinery working life prediction method utilizing

reliability and condition-monitoring data. *Proc. Instn Mech. Engrs, Part E: J. Process Mechanical Engineering*, 2000, **214**, 109–122.
11 **Sutherland, H., Repoff, T., House, M.,** and **Flickinger, G.** Prognostics, a new look at statistical life prediction for condition-based maintenance. *IEEE Aerospace Conf. Proc.*, 2003, **7**, 3131–3136.
12 **Frelicot, C.** and **Dubuisson, B.** K-step ahead prediction in fuzzy decision space-application to prognosis. In Proceedings of IEEE International Conference on Fuzzy Systems, San Diego, USA, 1992, pp. 669–676.
13 **Jun, M. J., Roumeliotis, S. I.,** and **Sukhatme, G. S.** State estimation of an autonomous helicopter using Kalman filtering. In Proceedings of the IEEE/RSJ International Conference on *Intelligent robots and systems*, Kyongju, South Korea, 1999, pp. 1346–1353.
14 **Cui, W.** A state-of-the-art review on fatigue life prediction methods for metal structures. *J. Mar. Sci. and Technol.*, 2002, **7**, 43–56.
15 **Wauer, J.** On the dynamics of cracked rotors: a literature survey. *Appl. Mechanics Rev.*, 1990, **43**, 13–17.
16 **Gasch, R.** A survey of the dynamic behavior of a simple rotating shaft with a transverse crack. *J. Sound and Vibr.*, 1993, **160**(2), 313–332.
17 **Edwards, S., Lees, A. W.,** and **Friswell, M. I.** Fault diagnosis of rotating machinery. *Shock and Vibr. Dig.*, 1998, **30**, 4–13.
18 **Mayes, I. W.** and **Davies, W. G. R.** The vibrational behavior of a rotating system containing a transverse crack. IMechE Conference on *Vibrations in rotating machinery*, Cambridge, UK, 1976, pp. 53–64.
19 **Davies, W. G. R.** and **Mayes, I. W.** The vibrational behavior of a multi-shaft, multi-bearing system in the presence of a propagating transverse crack. *J. Vibr., Acoustics, Stress, and Reliability in Des.*, 1984, **106**, 146–153.
20 **Rychlik, I.** A new definition of the rainflow cycle counting method. *Int. J. Fatigue*, 1987, **9**(2), 119–121.
21 **Baldwin, J. D.** and **Thacker, J. G.** A strain-based fatigue reliability analysis method. *J. Mech. Des.*, 1995, **117**(2), 229–234.
22 **Miner, M. A.** Cumulative damage in fatigue. *Trans. ASME, J. Appl. Mechanics*, 1945, **67**, 159–164.
23 **Murakami, Y., Harada, S., Endo, T., Tani-Ishi, H.,** and **Fukushima, Y.** Correlations among growth law of small cracks, low-cycle fatigue law and applicability of Miner's rule. *Engng Fracture Mechanics*, 1983, **18**(5), 909–924.
24 **Miller, K. J.** Materials science perspective of metal fatigue resistance. *Mater. Sci. Technol.*, 1993, **9**(6), 453–462.
25 **Paris, P.** and **Erdogan, F.** A critical analysis of crack propagation laws. *Trans. ASME, J. Basic Engng, Ser. D*, 1963, **85**(4), 528–534.
26 **Paris, P., Gomez, M. P.,** and **Anderson, W. E.** A rational analytical theory of fatigue. *Trend Engng*, 1961, **13**(1), 9–14.
27 **Gilbert, C., Dauskardt, R.,** and **Ritchie, R.** Microstructural mechanisms of cyclic fatigue-crack propagation in grain-bridging ceramics. *Ceram. Int.*, 1997, **23**, 413–418.
28 **Drucker, D.** and **Palgen, L.** On the stress–strain relations suitable for cyclic and other loading. *Trans. ASME, J. Appl. Mechanics*, 1981, **48**, 479–485.
29 **Needleman, A.** A continuum model for void nucleation by inclusion debonding. *Trans. ASME, J. Appl. Mechanics*, 1987, **54**, 525–531.
30 **Xu, G., Argon, A.,** and **Ortiz, M.** Nucleation of dislocations from crack tips under mixed-modes of loading–implications for brittle against ductile behavior of crystals. *Philosophical Mag.*, 1995, **72**, 415–451.
31 **Sadananda, K., Vasudevan, A. K., Holtz, R. L.,** and **Lee, E. U.** Analysis of overload effects and related phenomena. *Int. J. Fatigue*, 1999, **21**, 233–246.
32 **Elber, W.** Fatigue crack closure under cycle tension. *Engng Fracture Mechanics*, 1970, **2**, 37–45.
33 **Singh, N., Khelawan, R.,** and **Mathur, G. N.** Effect of stress ratio and frequency on fatigue crack growth rate of 2618 aluminum alloy silicon carbide metal matrix composite. *Bull. Mater. Sci.*, 2001, **24**(2), 169–171.
34 **Walker, K.** The effect of stress ratio during crack propagation and fatigue for 2024-T3 and 7075-T6 aluminum. ASTM STP 462, pp. 1–14, .
35 **Dowling, N. E.** *Mechanical behavior of materials: engineering methods for deformation, fracture, and fatigue*, 2006 (Prentice Hall, Upper Saddle River, New Jersey).
36 **Gere, J. M.** and **Timoshenko, S. P.** *Mechanics of materials*, 1984 (Chapman and Hall, London).
37 **Jiang, F., Rohatgi, A., Vecchio, K. S.,** and **Cheney, J. L.** Analysis of dynamic responses for pre-cracked three-point bend specimen. *Int. J. Fracture*, 2004, **127**(2), 147–165.
38 **Jiang, F., Rohatgi, A., Vecchio, K. S.,** and **Adharapurapu, R. R.** Crack length calculation for bend specimens under static and dynamic loading. *Engng Fracture Mechanics*, 2004, **71**, 1971–1985.
39 **Bakker, A.** Compatible compliance and stress intensity expressions for the standard three-point bend specimen. *Fatigue and Fracture Engng Mater. and Struct.*, 1990, **13**(2), 145–154.
40 **Saavedra, P. N.** and **Cuitino, L. A.** Vibration analysis of rotor for crack identification. *J. Vibr. and Control*, 2002, **8**(1), 51–67.
41 **Wu, S.**Development and validation of machinery health monitoring diagnostic, and prognostic methodology and tools. PhD dissertation, Department of Mechanical and Industrial Engineering, University of Illinois at Chicago, Chicago, Illinois, .
42 **Ross, S. M.** *Introductory statistics*, 2005 (Academic Press, Boston, Massachusetts).

APPENDIX 1

Notation

a_0	initial crack size
a_f	defined failure size
b_k	participation factors of individual harmonics
C, α, γ	material-, circumstance-, and load-related parameters
da/dN	crack growth rate
d	damping

D	dimensionless damping of the system $= d/2m\omega_0$
E	young's modulus of elasticity
G'	shear modulus
I	moment of inertia of the intact rotating shaft
L	span of the shaft
m	mass of the shaft
N_i	number of cycles to failure of a constant torque τ
N_T	remaining cycles
R	stress ratio
s_0	bending stiffness of an intact shaft
SO_{crack}	shaft order due to cracking
SO_{norm}	shaft order under normal condition
SO_ε	shaft order due to eccentricity
SO_1	first shaft order
SO_2	second shaft order
SO_3	third shaft order
$[SO_1^*]_{\text{cracked}}$	threshold of first shaft order
$\Im-\xi$	rotating coordinates
$w_{z,\text{stat}}$	weight deflection
y–z	inertial coordinates
Y	geometry factor depending on the loading and configuration
β	angle between ε and the centre of the crack
ΔK	amplitude of stress intensity factor
$\Delta r(t)$	forced excitation
$\Delta r_{\text{crack}}(t)$	forced excitation due to cracking
$\Delta r_\varepsilon(t)$	forced excitation due to eccentricity
Δs	change in stiffness
Δs_ξ	change in stiffness on direction ξ
$\Delta\sigma$	difference between maximum stress and minimum stress
ε	eccentricity
η	relative excitation frequency of the system $= \Omega/\omega_0$
λ	a constant depending on the rotational speed and torque
ν	poisson's ratio
σ	load
$\sigma_{\text{endurance}}$	endurance limit
τ	torque
φ	damage level of the shaft
Φ	diameter of shaft
ω_0	eigenfrequency of the rotor $= \sqrt{s_0/m}$
Ω	rotational frequency

APPENDIX 2

Computation of shaft orders

Let $\Gamma(f)$ be the shaft order of the vibration excitation due to imbalance and cracking. By applying a Fourier transform to equation (10), $\Gamma(f)$ can be obtained as

$$\begin{aligned}\Gamma(f) &= \int_{-T/2}^{+T/2}\left\{\varepsilon\left(\frac{\eta^2}{1-\eta^2+2\mathrm{j}D\eta}\right)\cdot\mathrm{e}^{\mathrm{j}(\Omega t+\beta)} + \lambda\sum_{k=-3}^{k=+3}\frac{b_k\mathrm{e}^{\mathrm{j}k\Omega t}}{1-k^2\eta^2+2\mathrm{j}Dk\eta}\right\}\cdot\mathrm{e}^{-2\pi\mathrm{j}ft}\,\mathrm{d}t \\ &= \int_{-T/2}^{+T/2}\left\{\varepsilon\left(\frac{\eta^2}{1-\eta^2+2\mathrm{j}D\eta}\right)\cdot\mathrm{e}^{\mathrm{j}\beta}\cdot\mathrm{e}^{(\mathrm{j}\Omega-2\pi f\mathrm{j})\cdot t} + \lambda\sum_{k=-3}^{k=+3}\frac{b_k}{1-k^2\eta^2+2\mathrm{j}Dk\eta}\cdot\mathrm{e}^{(k\Omega-2\pi f)\mathrm{j}t}\right\}\mathrm{d}t \\ &= \varepsilon\left\{\frac{\eta^2}{1-\eta^2+2\mathrm{j}D\eta}\right\}\mathrm{e}^{\mathrm{j}\beta}\cdot\frac{\mathrm{e}^{\mathrm{j}T(\Omega/2-\pi f)}-\mathrm{e}^{\mathrm{j}T(\pi f-\Omega/2)}}{2\pi f-\Omega}\mathrm{j} + \lambda\sum_{k=-3}^{k=+3}\frac{b_k}{1-k^2\eta^2+2\mathrm{j}Dk\eta}\cdot\frac{\mathrm{e}^{(k\Omega-2\pi f)\mathrm{j}T}}{k\Omega-2\pi f}\end{aligned} \tag{12}$$

where T is the duration of the vibration. Thus, first shaft order SO_1 can be obtained when f is set equal to rotational frequency Ω

$$\begin{aligned}SO_1 = \Gamma(\Omega) &= \varepsilon\left\{\frac{\eta^2}{1-\eta^2+2\mathrm{j}D\eta}\right\} \times\frac{\mathrm{e}^{\mathrm{j}[\beta+T\Omega(1/2-\pi)]}-\mathrm{e}^{\mathrm{j}[\beta+T\Omega(\pi-1/2)]}}{\Omega(2\pi-1)}\mathrm{j} \\ &\quad+\lambda\sum_{k=-3}^{k=+3}\frac{b_k}{1-k^2\eta^2+2\mathrm{j}Dk\eta}\cdot\frac{\mathrm{e}^{(k-2\pi)\mathrm{j}\Omega T}}{\Omega(k-2\pi)}\end{aligned} \tag{13}$$

In similar fashion, the second and third shaft order can be computed as $SO_2 = \Gamma(2\Omega)$ and $SO_3 = \Gamma(3\Omega)$ respectively.

APPENDIX 3

Computation of constant λ using shaft order 1

Knowing SO_1, from equation (13), the constant term λ can be computed as

$$\lambda = \left|\frac{SO_1-\varepsilon\{\eta^2/(1-\eta^2+2\mathrm{j}D\eta)\}\{(\mathrm{e}^{\mathrm{j}[\beta+T\Omega(1/2-\pi)]}-\mathrm{e}^{\mathrm{j}[\beta+T\Omega(\pi-1/2)]})/[\,\Omega(2\pi-1)]\}\mathrm{j}}{\sum_{k=-3}^{k=+3}\frac{b_k}{(1-k^2\eta^2+2\mathrm{j}Dk\eta)}\cdot\frac{\mathrm{e}^{(k-2\pi)\mathrm{j}\Omega T}}{\Omega(k-2\pi)}}\right| \tag{14}$$

Journal of
Nanoscience and Nanotechnology
Vol. 8, 6065–6074, 2008

Large Scale Combustion Synthesis of Single-Walled Carbon Nanotubes and Their Characterization

Henning Richter[1,*], Meri Treska[2], Jack B. Howard[1], John Z. Wen[3], Sebastien B. Thomasson[2], Arthur A. Reading[2], Paula M. Jardim[2,4], and John B. Vander Sande[2]

[1]Nano-C, Inc., 33 Southwest Park, Westwood, MA 02090, USA
[2]Department of Materials Science and Engineering and
[3]Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
[4]Department of Materials Science and Metallurgy, Pontifical Catholic University, Rio de Janeiro, 22453-900, Brazil

Since its invention in 1991, premixed combustion synthesis of fullerenic materials has been established as the major industrial process for manufacturing of these materials. Large-scale production of fullerenes such as C_{60}, C_{70} and C_{84} has been implemented. More recently, combustion technology has been extended to the targeted synthesis of single-walled carbon nanotubes (SWCNT). Addition of catalyst precursor and operation at well-controlled fuel-rich but non-sooting conditions are required. Extensive parametric studies have allowed for the optimization of the formation of high-quality SWCNT. Purification techniques previously reported in the literature have been adjusted and used successfully for the nearly complete removal of metal and metal oxide. Material has been characterized using Raman spectroscopy, scanning (SEM) and transmission electron microscopy (TEM), scanning transmission electron microscopy (STEM), atomic force microscopy (AFM), X-ray diffraction (XRD), and thermogravimetric analysis (TGA). Correlations between process conditions and nanotube properties such as length have been established. Product reproducibility and process scalability of the combustion process have been demonstrated. Sample preparation was found to affect significantly the apparent characteristics of nanotubes as seen in electron microscopy images.

Keywords: SEM, TEM, AFM, XRD, TGA.

RESEARCH ARTICLE

1. INTRODUCTION

Availability of homogenous lots of single-walled carbon nanotubes (SWCNT) with well-defined and reproducible characteristics in sufficiently large quantities and at low price is a condition for the successful commercial development of promising applications ranging from selective gas sensors,[1–3] the reinforcement of polymers,[4–5] drug delivery,[6–7] to electronics[8–11] and solar cells.[12] While first products are in the process of reaching the market, large-scale commercialization could be prohibited by increasing concerns over currently available SWCNT availability and quality.[13]

One of the purposes of the present paper is to demonstrate the potential of combustion technology for the cost-efficient large-scale manufacturing of SWCNT. Since the initial discovery of SWCNT in a plasma discharge in the presence of an iron catalyst,[14] a large spectrum of different methods allowing for the synthesis of SWCNT including laser and chemical vapor deposition techniques has been investigated using in most cases supported or floating catalysts.[15] While details can differ significantly, nearly all approaches are based on the transformation of carbon, for instance in the form of a hydrocarbon, to SWCNT in the presence of heat and a metal catalyst. While the combustion process used in this work is based on similar principles, the absence of an additional external energy source together with a well-defined and controllable set of operating conditions points to its particular scalability. The suitability of combustion for the large-scale manufacturing of carbonaceous nanomaterials has been convincingly demonstrated for carbon black[16] and, more recently, fullerenes.[17–21] Analysis of material generated by means of premixed combustion revealed the presence of carbonaceous nanostructures[22,23] and isolated single- and multi-walled carbon nanotubes in the absence[24–27] and presence[28] of explicitly added catalyst. Growth of multi-walled carbon nanotubes (MWCNT) on metal substrates was reported in

*Author to whom correspondence should be addressed.

diffusion,[29, 30] opposed-flow diffusion[31] but also premixed flames.[32] The effect of electric fields on alignment and growth rate of MWCNT in opposed-flow diffusion flames has been investigated.[33]

Synthesis of SWCNT in pyrolysis flames has been studied extensively by Vander Wal and co-workers.[34–36] In this approach, catalyst particles were generated, e.g., by nebulizing iron salt solutions added to a fuel such as CO/H_2. This reactive mixture was introduced into a fuel-rich acetylene flame via a co-centric tube surrounded by the sintered metal plate of a premixed flame burner. Variation of mixture composition and flow rates allowed for the optimization of SWCNT formation. In addition, formation of SWCNT in diffusion flames has been observed using a floating catalyst formed from a metallocene.[37]

The SWCNT-synthesis approach described in the present contribution is based on the addition of a catalyst precursor such as iron pentacarbonyl ($Fe(CO)_5$) to the fresh gas mixture prior to the stabilization of a premixed flat flame. It is the result of the further development of the work of Height et al.[38–40] In a detailed analysis of premixed acetylene/oxygen/argon/$Fe(CO)_5$ low pressure flames, operating conditions suitable for the formation of SWCNT were optimized.[38, 39] Different fuel-to-oxygen ratios and $Fe(CO)_5$ concentrations were investigated. A nanotube formation window of fuel-to-oxygen ratios was identified. While, relative to stoichiometric conditions, an excess of fuel (i.e., "carbon") is necessary to enable inception of nanotubes, soot-like structures are formed at too high fuel-to-oxygen ratios. In the case of insufficient fuel supply, metal and metal oxide particles without nanotubes attached have been observed using transmission electron microscopy (TEM). The most pronounced effect of $Fe(CO)_5$ concentration was found to be on the particle size distribution and the shape of the metallic particles. The quantity of condensed material increased dramatically with the $Fe(CO)_5$ concentration whereas nanotubes appeared to be cleaner at lower concentrations.

2. EXPERIMENTAL DETAILS

At Nano-C, combustion synthesis of SWCNT was further developed in view of its use for large-scale manufacture. In-line filtration allowing for continuous operation and product recovery was implemented using commercial metallic filter cartridges. Major improvements of the SWCNT yields and their quality have been achieved by using natural gas or methane as fuel. Consistent with the, compared to acetylene, significantly lower sooting tendency of small saturated hydrocarbons such as methane,[41] operation at higher fuel-to-oxygen ratios, i.e., supplying more "carbon" for SWCNT growth without competing soot formation was found to be possible.

Meaningful characterization is essential for the identification of SWCNT suitable for specific applications and the quality control manufacturing processes. For instance, reproducibility and well-defined correlations between process parameters and materials characteristics need to be established. In the present work, a large range of operating conditions has been investigated and samples collected. While first assessments of the presence of SWCNT have been conducted by means of Raman spectroscopy, more detailed information was obtained using mostly SEM and TEM but also STEM and AFM. Sample preparation, such as the use and extent of sonication, was found to have a significant impact on the resulting images in the SEM, TEM and AFM observations. The development of well-defined characterization procedures has also been essential for the investigation of the SWCNT formation mechanism. Using the same reactor, samples withdrawn iso-kinetically at different heights above the burner have been analyzed.[42]

In the present work, operation conditions were optimized for the formation of SWCNT at different pressures using Raman intensities as criterion. Purification techniques previously reported in the literature[43–45] have been adjusted. Remaining metal or metal oxide particles (from catalyst) have been nearly completely removed using a HCl followed by HNO_3 treatment or a sequence of oxidation and HCl-reflux steps. Both optimized as-produced material and purified SWCNT have been investigated in some detail by means of SEM and TEM using a range of sample preparation techniques. To examine the purity of the products and the reproducibility of the process, TGA and XRD were used.

Using a modified laboratory scale reactor, a production capability of 5 to 10 kg of as-produced SWCNT has been established at Nano-C.[46] The process is easily scalable, as demonstrated in fullerenes ($C_{60}, C_{70}, \ldots, C_{84}$) manufacturing. Equipment similar to the currently used SWCNT reactor has been scaled up to a capacity of 1 to 2 tons/year.

2.1. Raman Spectroscopy

Resonance Raman spectroscopy has been demonstrated to be a fast and selective method for the identification and first characterization of SWCNT.[47–49] Major features are the radial breathing mode (RBM), the tangential mode (G-band) and the disorder-induced band (D-band). RBM, usually appearing between $120\ cm^{-1} < \omega_{RBM} < 270\ cm^{-1}$, correspond to the atomic vibration of the C atoms in the radial direction and direct correlations with SWCNT diameters have been established.[47–49] The G-band, a characteristic multi-peak feature around $1580\ cm^{-1}$, corresponds to atomic displacements along the tube axis as well as the circumferential direction. Simultaneous observation of RBM and G-band provides strong evidence for the presence of SWCNT. The D-band, occurring around $1350\ cm^{-1}$, reflects the presence of impurities or other symmetry-breaking defects such as amorphous carbon.[50] In the present work, material synthesized and

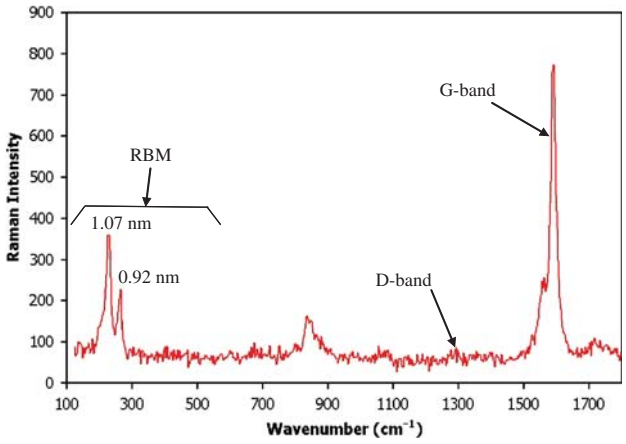


Fig. 1. Raman spectrum of typical as-produced material measured at 784.87 nm.

collected under well defined conditions was investigated with a Dimension-P2 Raman system (Lambda Solutions, Waltham, MA) using a exciting wavelength of 784.87 nm and approximately 10 mW power. A laser beam with a diameter of about 200 micrometer was directed without microscope to the samples from a distance of ≈1 cm. Usually, 5 s exposure time and integration over 5 spectra were applied. Peak heights of RBM- and G-band were optimized by fine-tuning the distance between the laser probe and the samples. In the process of identifying best operating conditions, methane-to-oxygen ratio, inert gas dilution, fresh gas velocity and iron pentacarbonyl concentration were varied systematically at different pressures between 50 and 400 torr. Samples with strong RBM- and G-bands as well as high G- to D-band ratios have been considered as being at or close to the optimized conditions and submitted to further analysis such as SEM and TGA, the latter allowing for a quantitative assessment of SWCNT abundance in given samples. All optimized flames have been fuel-rich and relatively close to the sooting limit while required iron pentacarbonyl concentrations decreased with increasing pressure. A typical Raman spectrum of as-produced material generated at optimized operating conditions measured at 784.87 nm is shown in Figure 1. The D-band was found to be barely detectable while RBM peaks at 229.6 and 265.5 cm^{-1} were identified. Using the relationship $\omega_{RBM} = 234/d_i + 10$ cm^{-1}, as suggested by Milnera et al.[51] for bundles of SWCNT, these peaks correspond to diameters of 1.07 and 0.92 nm, respectively. However, due the strong dependence of Raman intensities on the resonance energies of the SWCNT present, such diameter distribution reflects only SWCNT resonating at 784.87 nm and is not representative for the investigated sample. For instance, a Raman spectrum of similar material measured at 647 nm gives a significantly different picture: RBM peaks corresponding to 1.30, 0.98 and 0.87 nm have been identified (Fig. 2).

2.2. Scanning Electron Microscopy (SEM)

SEM of as-produced and purified SWCNT has been conducted in the Center of Materials Science and Engineering (CMSE) at MIT using a FEI/Philips XL30 FEG ESEM instrument. SEM is a relatively fast technique allowing for a large range of sample preparation techniques and sufficient for the visualization of SWCNT bundles but not

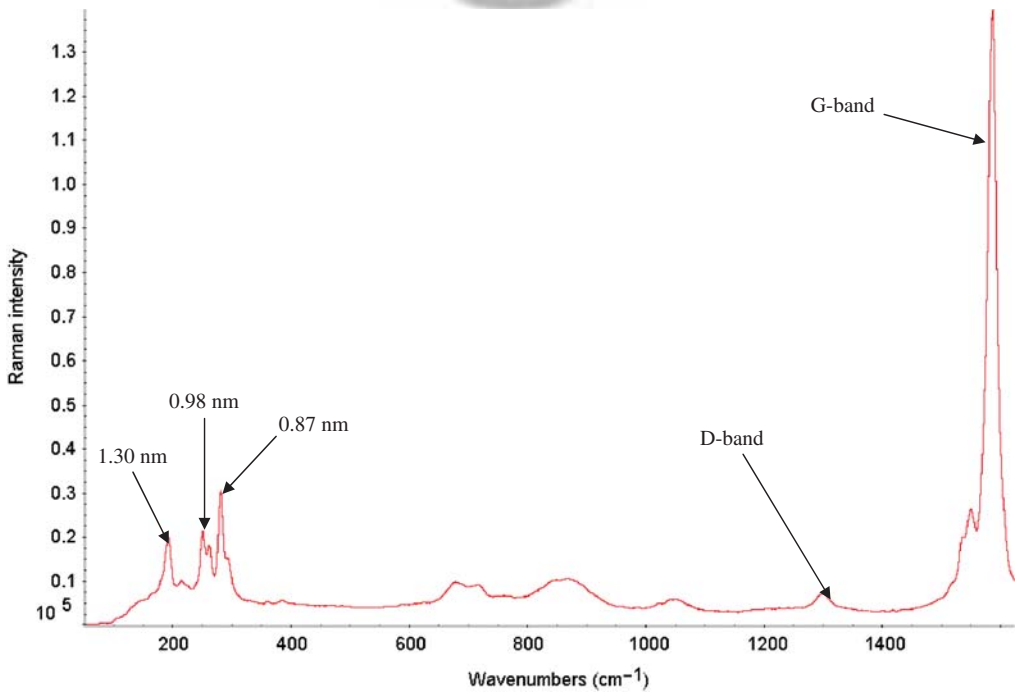
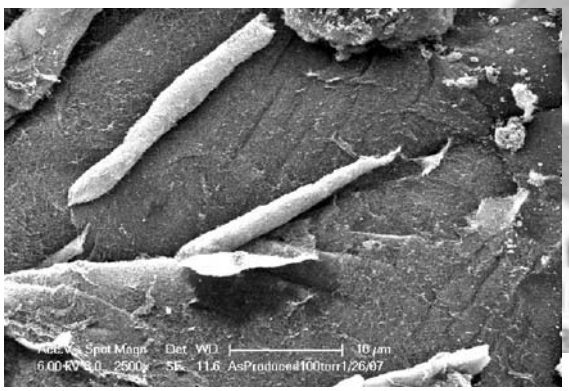


Fig. 2. Raman spectrum of typical as-produced material measured at 647 nm.

individual tubes. In the present work, two major sample preparation techniques have been used routinely: (a) deposition of as-received material on a carbon film (minimized sample preparation) and (b) dispersion in ethanol followed by tip-mode sonication (30 watt, 20 kHz, 1–2 min) using a Vibra cell™ instrument (Sonics and Materials, Danbury, CT) and followed by deposition on a holey carbon grid. In addition, the effect of sonication in other solvents such as chloroform has been assessed. Figure 3 shows micrographs of as-produced material after minimized sample preparation at magnifications of 2500 (top) and 50000 (bottom). Micron markers are also in the figures. While sheet-like structures can be seen at a magnification of 2500, a network of SWCNT bundles is visible at the higher magnification. Even at highest magnifications, the resolution of SEM is not sufficient to image individual SWCNT; other techniques, particularly TEM and AFM, are required. Another interesting feature of the images in Figure 3 is the under-representation of metal and metal oxide particles. While clearly identified and even quantified by means of TEM, XRD and TGA, they are barely visible in SEM using the described minimized sample preparation. Figure 4 provides a first impression of the effect of sample preparation on resulting SEM images. A sample of the same batch of starting material as that imaged in Figure 3 was sonicated in ethanol and deposited on a holey carbon grid. Wrapping of SWCNT bundles around the holey carbon grid can be observed at a magnification of 15000 (Fig. 4, top) whereas a rather inhomogeneous distribution of SWCNT can be seen at higher magnifications (bottom).

(a) ×2500



(b) ×50000

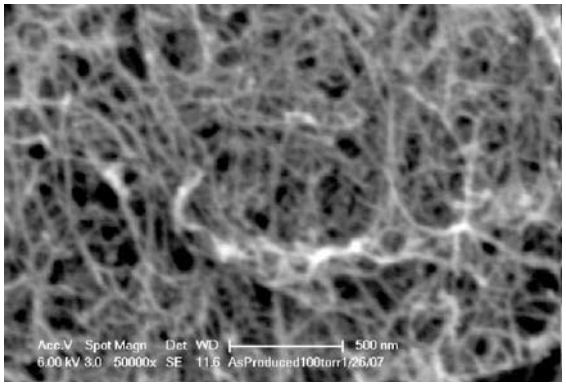
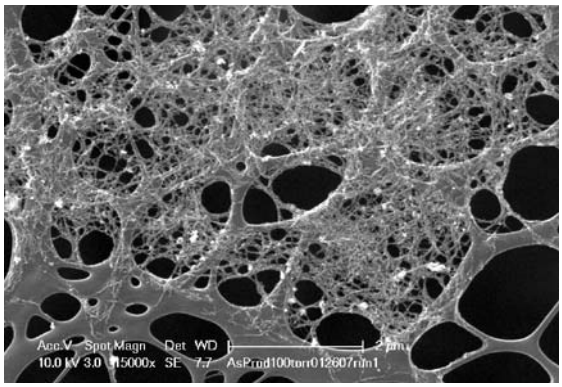


Fig. 3. Scanning electron microscopy (SEM) of as-produced material deposited on a carbon film. (a) magnification of 2500, (b) magnification of 50000. Magnification is also shown in each image through a micron marker. These images represent typical results for all sample areas.

(a) ×15000



(b) ×35000

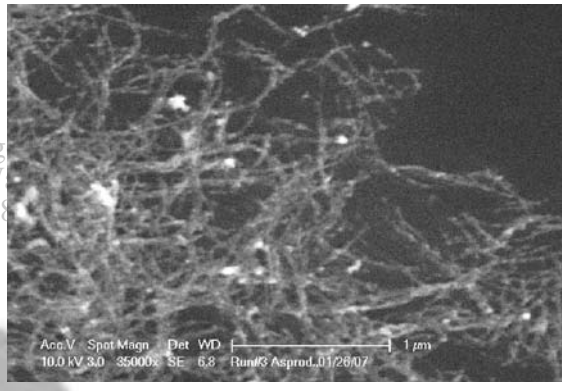


Fig. 4. Scanning electron microscopy (SEM) of as-produced material dispersed in ethanol, sonicated and deposited on a holey carbon grid. Magnification: (a) 15000 and (b) 35000. Magnification is also shown in each image through a micron marker.

A general characteristic of purified SWCNT is the enhanced difficulties to image the bundles present due to strong inter-molecular forces. In the event of dispersion, increased sonication times for the efficient break up of SWCNT clusters are required. Figure 5 shows SEM images of purified SWCNT at different magnifications (2500, 20000, 65000, 100000) using minimized sample preparation. Typically, "rock-like" structures are seen (Fig. 5(a)) but at higher magnifications a dense network of entangled bundles of SWCNT becomes visible. At edges of clusters, free-standing bundles of SWCNT could be identified. In addition to the method of sample preparation used, also the substrate has a significant impact on the resulting image. Purified SWCNT have been sonicated in ethanol, i-propanol or toluene and deposited either on a flat carbon grid or a holey carbon grid. While SEM imaging of the sample on the flat carbon film shows relatively densely packed agglomerates of the SWCNT bundles, much more homogeneously dispersed bundles could be seen on the

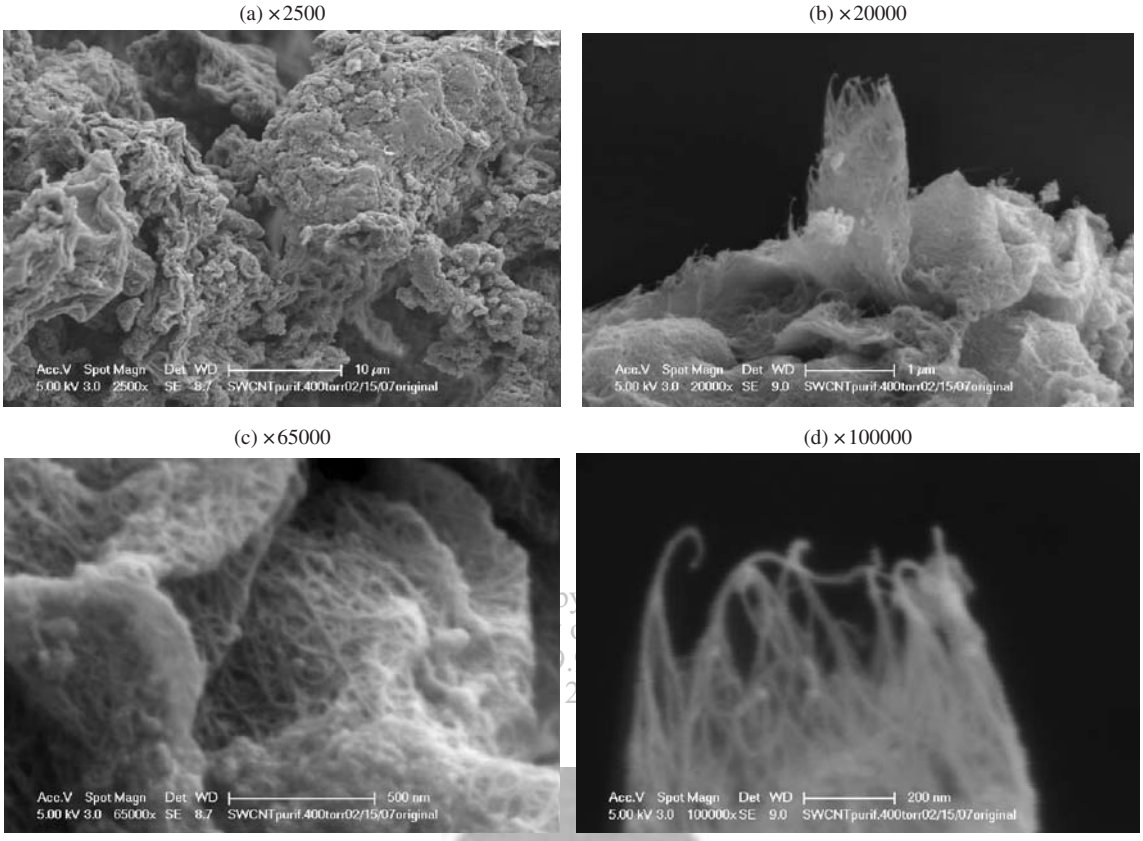


Fig. 5. Scanning electron microscopy (SEM) of purified SWCNT deposited on a carbon film. Magnification: (a) 2500, (b) 20000, (c) 65000 and (d) 100000. Magnification is also shown in each image through a micron marker.

holey carbon grid (Fig. 6). The high flexibility of the SWCNT bundles is emphasized by their wrapping around the frame of the holey carbon grid (Fig. 6(b)).

Ring structures of carbon nanotubes,[52, 53] biological filaments[53] and amorphous carbon nanomaterials[54] have been reported previously. In the present work, rings of SWCNT bundles, with diameters ranging from approximately 1.5 to 6 μm, have been formed after sonication in chloroform and deposition on a silicon wafer (Fig. 7). This is consistent with the work of Martel et al.[52] describing ring formation from straight SWCNT after short sonication in 1,2-dichloroethane. Radii (300–400 nm) are typically smaller than in the present work, probably due to differences in the length of the initial SWCNT. Martel et al.[52] explained ring formation as a balance between tube–tube van der Waals adhesion and the strain energy resulting from the coiling-induced curvature. Ring formation is believed to be kinetically controlled with bubble cavitation, generated by ultrasonic irradiation, providing the necessary energy.

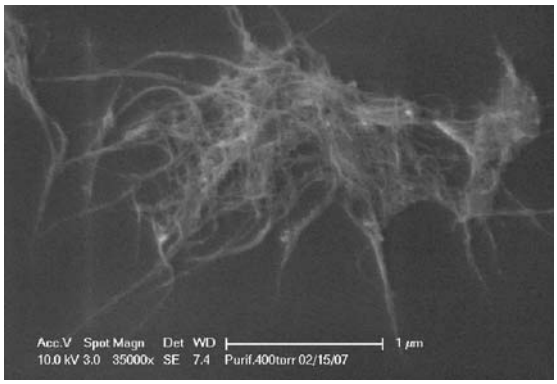
Length of SWCNT has been found to be a critical characteristic for many applications. While dispersion can be expected to be more challenging with increasing length, an increase of the conductivity of SWCNT networks corresponding to length$^{1.46}$ for bundles of the same diameter has been determined.[55] A standard procedure for the assessment of the length of the SWCNT bundles has been developed in the present work. As-produced or purified SWCNT were dispersed in ethanol and sonicated (tip mode, 30 watt, 20 kHz). Sonication times were usually 1–2 min for as-produced material and up to 10 min for purified SWCNT. Prior to deposition on holey carbon grids, the dispersion was diluted with additional ethanol. During SEM imaging at magnifications of typically 10000 to 20000, SWCNT bundles were identified and lengths measured manually using a feature of the instrument's control software (Fig. 8). Mean values and standard deviations have been determined. Figure 8 shows a sample of purified SWCNT for which an average bundle length of 1.5 ± 0.5 μm has been obtained based on 42 bundles ranging from 0.7 to 2.4 μm. Since beginning and end of bundles are not always visible and bending has not been taken into account, reported values represent rather conservative lower limits.

2.3. Transmission Electron Microscopy (TEM) and Atomic Force Microscopy (AFM)

While the presence of SWCNT has been confirmed unambiguously by means of Raman spectroscopy, SEM alone is not sufficient for their visualization. SEM provides a highly valuable overview of the sample appearance but

RESEARCH ARTICLE

(a) ×35000



(b) ×65000

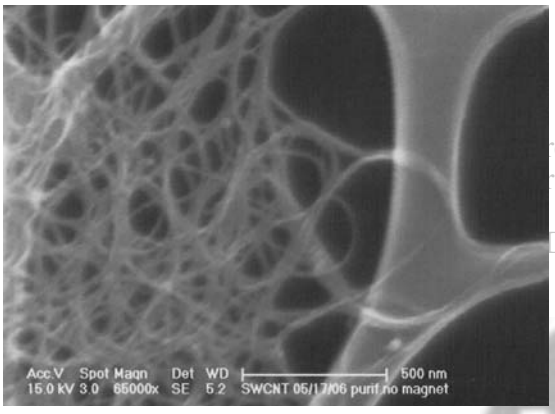


Fig. 6. Scanning electron microscopy (SEM) of purified SWCNT sonicated in ethanol. (a) deposited on carbon film, ×35000, (b) deposited on a holey carbon grid, ×65000. Magnification is also shown in each image through a micron marker.

resolution is too poor to show individual SWCNT. The use of complementary techniques allowing for higher magnifications, particularly TEM (using a Jeol 2010 instrument operating at 200 keV) but also AFM, is therefore necessary.

In order to establish the link between SEM and TEM, the same sample of purified SWCNT, deposited on a holey

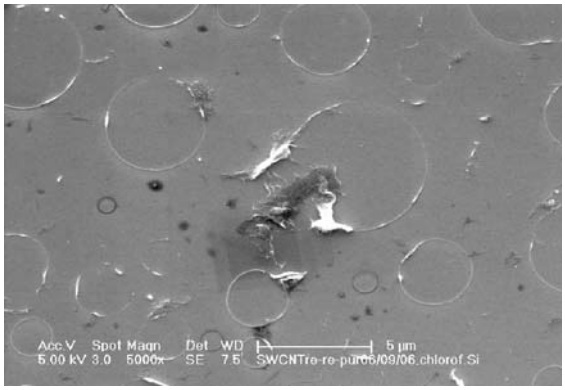


Fig. 7. Scanning electron microscopy (SEM) of purified SWCNT sonicated in chloroform. Deposited on silicon wafer. Magnification: ×5000. Magnification is also shown through a micron marker. Note the bending and "ring" formation of bundles of nanotubes.

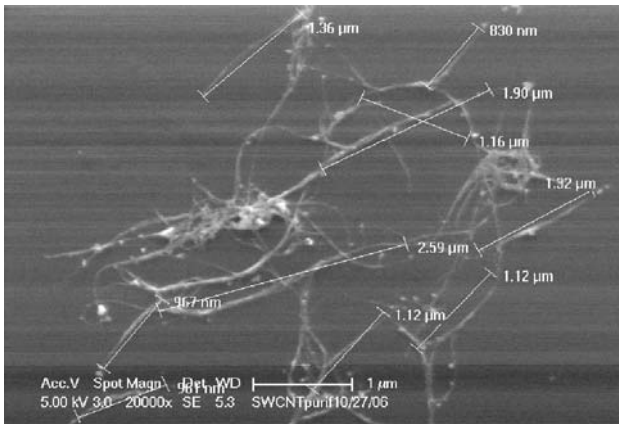
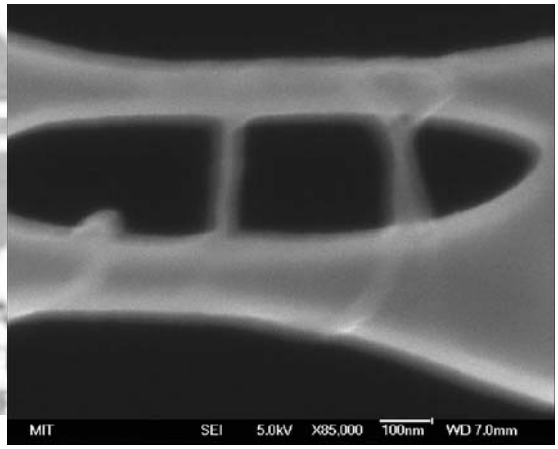


Fig. 8. Determination of length distribution of SWCNT bundles. Sonicated in ethanol, diluted and deposited on a holey carbon grid.

carbon grid after sonication in ethanol, was imaged by means of both techniques, using magnifications of approximately 85000. SEM and low-magnification TEM were taken from the same area of the grid and a striking similarity could be observed (Fig. 9). In both cases, the tubes have wrapped around the web in the lacey carbon

(a)



(b)

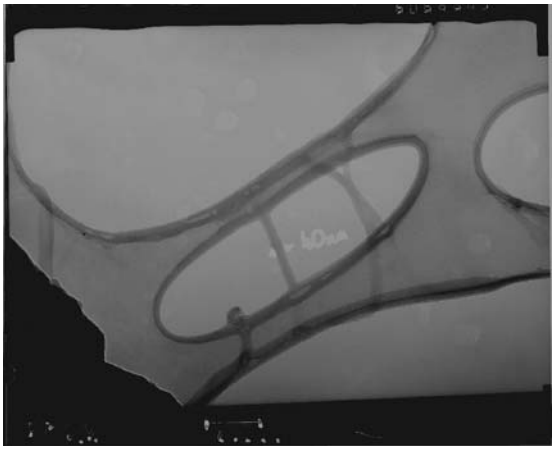


Fig. 9. SEM (a) and low-magnification TEM (b) of the same sample on a holey carbon grid after deposition of purified SWCNT sonicated in ethanol. Figure (b) is 0.73 the magnification of (a).

grid. This "bridge" between images obtained by SEM and TEM gives us confidence to extend the SEM method as a routine method for observing nanotube structures. Significantly higher TEM magnifications, of up to 300000, allowed for the visualization of individual SWCNT. Dispersion in ethanol, followed by sonication and deposition on holey carbon grids has been used for all TEM images. Figure 10 shows a TEM image of as produced material with a single SWCNT shown in the insert. In comparison to SEM micrographs of similar material after using the same sample preparation (Fig. 4) particles are much more visible, indicating the effect of the imaging technique. In addition to bundles of SWCNT, some spherical, fullerene-like structures appear to be present on the surface. Their identity will be assessed in some more detail below. A TEM image of purified SWCNT is shown in Figure 11. There are mainly densely packed bundles of SWCNT but measurements of diameters of some single tubes have been possible. Comparison with Figure 10 seems to indicate generally larger diameters of purified SWCNT, possibly due to partially selective destruction of smaller ones in the purification process. However, only very small amounts were needed for sample preparation and quantitative assessments based on a limited number of images and SWCNT cannot be considered as representative. The impression of the presence of less smooth surfaces in the case of purified SWCNT may reflect defects induced by the purification process. Less aggressive purification procedures are under development. De-bundling of SWCNT using surfactants such as sodium dodecyl sulfate (SDS),[56,57] sodium dodecylbenzene sulfonate (NaDDBS)[57] or sodium cholate,[58] also necessary for detailed characterization by means of optical spectroscopy,[56,58] is currently under investigation.

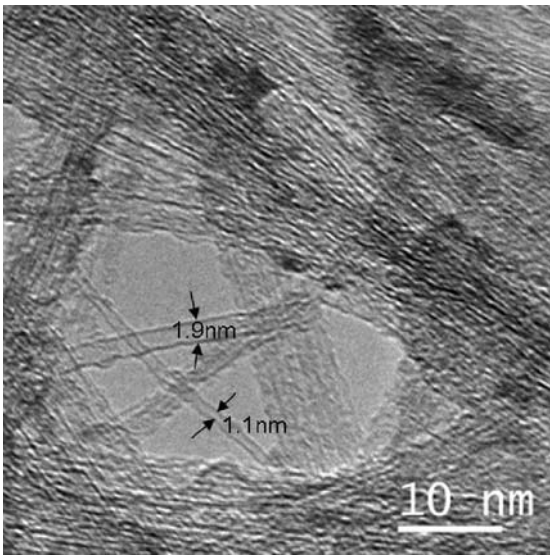


Fig. 11. TEM of purified SWCNT. Sonicated in ethanol followed by deposition on a holey carbon grid. Magnification is also shown through a micron marker.

AFM of purified SWCNT deposited on a silicon wafer after sonication in ethanol and dilution has been conducted using a DIMENSION D 3000 instrument with a Nanoscope IIIa controller (Fig. 12). The morphology of individual SWCNT could be assessed in three dimensions and with good contrast. Similar to TEM images of

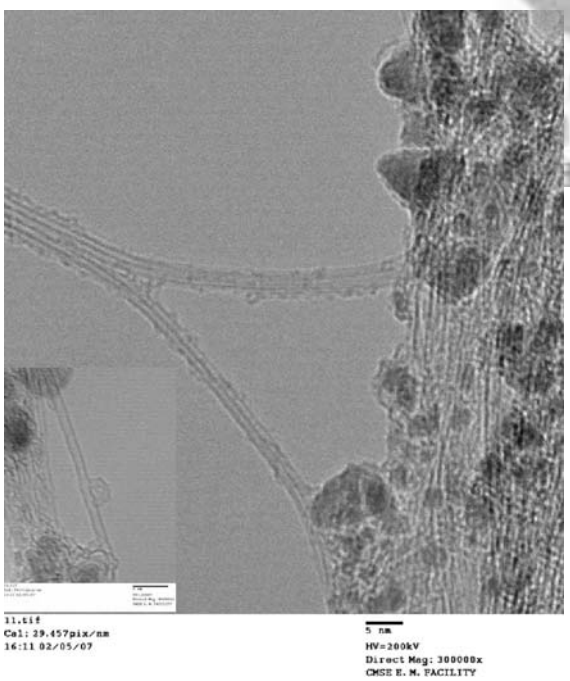


Fig. 10. TEM of as-produced SWCNT. Sonicated in ethanol followed by deposition on a holey carbon grid. Magnification is shown through a micron marker.

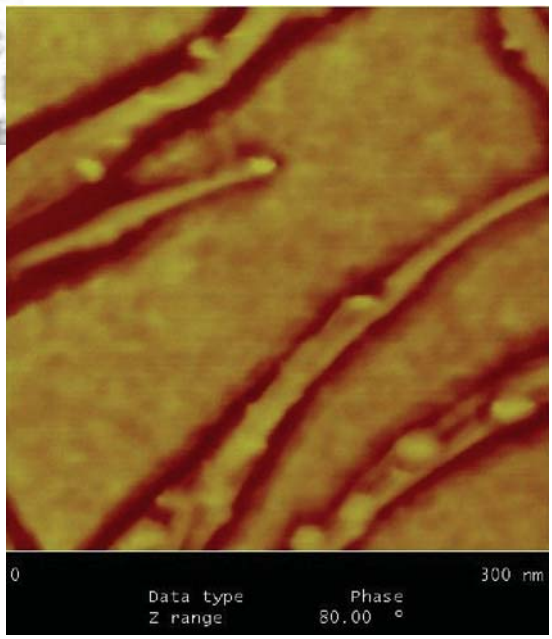


Fig. 12. AFM of purified SWCNT. Sonicated in ethanol followed by deposition on a silicon wafer. Magnification is shown through a micron marker. The entire width of the image is equal to 300 nm.

RESEARCH ARTICLE

as-produced (Fig. 10) and purified (Fig. 11) SWCNT, spherical structures have been observed. Their identity has been assessed by means of scanning transmission electron microscopy (STEM).

2.4. Scanning Transmission Electron Microscopy (STEM)

Additional techniques are required for the characterization of material other than SWCNT, present in as-produced and—to a much lesser extent—purified material. Electron diffraction, a feature built in to transmission electron microscopes, indicates the presence of elemental iron particles in purified SWCNT and of both elemental iron and maghemite, Fe_2O_3, in as-produced material. Although dark-field microscopy can be used, more detailed assessment such as spatial mapping of these particles requires the use of scanning transmission electron microscopy (STEM). Coupling with energy-dispersive X-ray spectroscopy allows for the determination of the composition directly correlated with the location in the sample. A VG-HB603 instrument, operated at 250 keV with a probe size of 1.5 nm FWHM (full width at half of maximum) has been used in the present work. Careful analysis of typical as-produced material revealed the presence of both iron oxide and iron particles whereas only elemental iron was identified in purified SWCNT. A STEM image of as-produced material is shown in Figure 13, together with X-ray spectra of two highlighted particles. The atomic ratio between iron and oxygen has not been quantified but additional information from electron diffraction and X-ray diffraction of bulk material (see below) provides a high level of confidence that there is maghemite (Fe_2O_3) present. We used the drift-correction capabilities of the Oxford Instruments Inca system installed on the STEM to ensure the spatial precision of the individual particle analyses. The points of the analysis were confirmed after the analysis by checking the

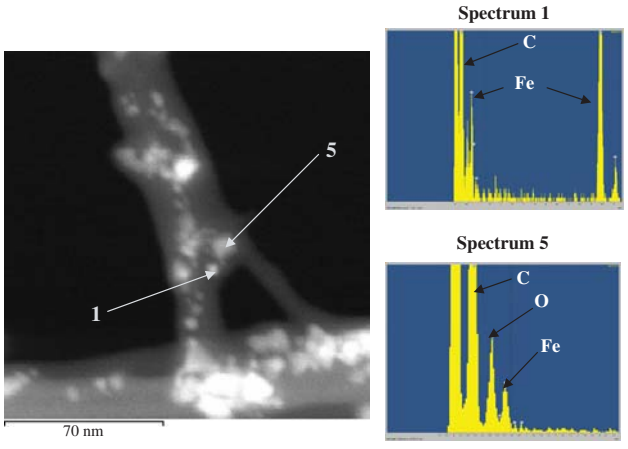


Fig. 13. STEM image of as-produced SWCNT. Sonicated in ethanol followed by deposition on a holey carbon grid. X-ray spectra measured at locations 1 and 5. Magnification is shown in the STEM image through a micron marker.

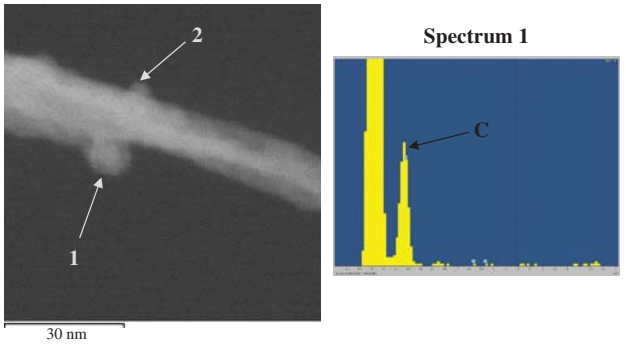


Fig. 14. STEM image of a bundle of purified SWCNT. Sonicated in ethanol followed by deposition on a holey carbon grid. X-ray spectrum measured at location 1.

contamination spot that was generally visible. Additional insight could be also gained concerning the identity of the spherical structures on the surface of some SWCNT bundles (Figs. 10 and 11). A STEM analysis of such structure, as shown in Figure 14, reveals the absence of any iron or oxygen and the exclusive presence of carbon, consistent with fullerene-like molecules.

2.5. Assessment of Purity

While electron diffraction and, particularly, STEM provide valuable information about the chemical identity of impurities of as-produced and purified material, these techniques are not suitable for the quantitative assessment of the purity of bulk samples, necessary for quality control. In the present work, both purity and reproducibility have been determined by means of thermogravimetric analysis (TGA) under air using a TGA i 1000 instrument (Instrument Specialists, Twin Lakes, WI). Heating rates of usually 5 or 7.5 K/min from room temperature to 900 °C have been applied. A typically TGA plot of as produced SWCNT is shown in Figure 15. Analysis of multiple batches produced at identical process conditions led to very similar, nearly overlapping plots, indicating a high degree of reproducibility.

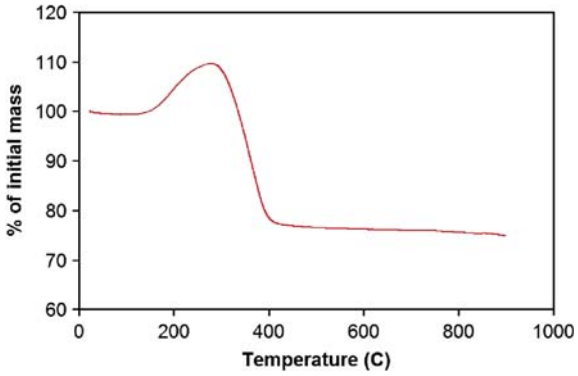


Fig. 15. Thermogravimetric analysis (TGA) under air of as-produced SWCNT. Heating rate: 5 K/min.

Quantification of the contents of carbonaceous material requires the knowledge of the composition of the metal phase in the initial sample in order to account for increase of mass by oxidation of elemental iron. Quantitative characterization of the metal phase has been conducted by means of Wide-Angle X-Ray Diffraction (XRD) using a Rigaku RU300 X-ray generator. Silicon (Si) has been added as an internal standard. XRD-patterns of both as-produced and purified SWCNT are given in Figure 16. Consistent with the electron diffraction and STEM results, both maghemite (Fe_2O_3) and elemental iron (Fe) have been identified in the as-produced material whereas only some elemental iron (Fe) was found to remain in purified SWCNT. Dissolution of elemental iron is likely to be more difficult because oxidation is required as a first step. Such oxidation can be carried out either by means of heating under air or the use of an oxidizing acid such as HNO_3. Oxidation can be prohibited by the coating of metal particles with carbonaceous material or their incorporation in SWCNT or their bundles. The possibility of preferential incorporation of iron (in contrast to iron oxide) in SWCNT and corresponding correlations with the SWCNT growth mechanism is currently under examination.

Using the internal standard, analysis of the XRD spectrum of the material for which the TGA plot is shown in Figure 15 led to a Fe-to-Fe_2O_3 weight ratio of 1.40. Assuming complete oxidation of elemental iron to Fe_2O_3 during the TGA run, weight fractions of 34.9% of Fe, 24.9% of Fe_2O_3 and 40.2% of carbonaceous material have been determined based on Figure 15. Consistent with the SEM and TEM results, the absence of a change of curvature (i.e., of a maximum of the first derivative) in the descending part of the plot, indicates the presence of only one major type of carbonaceous material, i.e., SWCNT. TGA runs of purified SWCNT showed a remaining mass of $\leq$5.5%. Assuming again complete oxidation to Fe_2O_3, this value corresponds to a purity of $\geq$96.1%. Higher degrees of purity could be achieved by a repetition of the purification procedure but also prolonged reflux times.

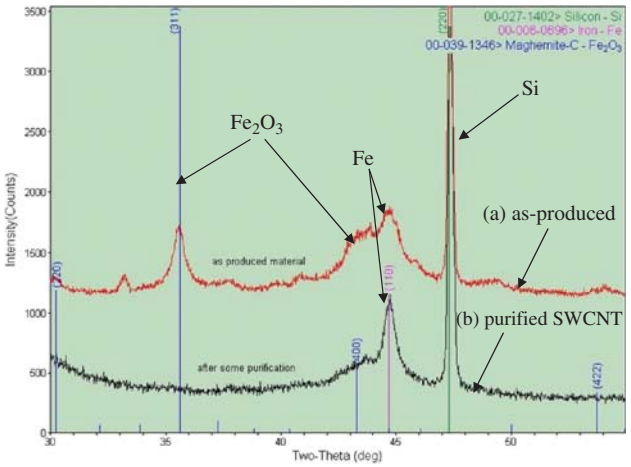


Fig. 16. X-ray diffraction (XRD) patterns of (a) as-produced and (b) purified SWCNT.

3. CONCLUSIONS

The efficiency of premixed combustion of natural gas or methane with a catalyst-precursor additive for the reproducible synthesis of SWCNT has been demonstrated. The process is exothermic, continuous and controlled with well-defined parameters. A procedure for the optimization of the operating conditions has been established. Raman spectroscopy has been found to be an effective and unambiguous screen for the presence of SWCNT. Peak intensities have been used for an initial assessment of the relative abundance of SWCNT in material collected on in-line filters. More detailed characterization was conducted using of SEM, TEM, AFM and STEM. While SEM allowed for the identification of general structural features of networks of bundles of SWCNT, higher magnifications, achieved with TEM and AFM, allowed for the visualization of individual tubes. The presence of SWCNT and the total absence of multi-walled carbon nanotubes were confirmed. Sample preparation and the structure of the substrate (e.g., carbon film vs. holey carbon grid) were found to have a significant impact on resulting images. Ring formation was observed after sonication in chloroform and deposition on a silicon wafer. Metal particles were more visible in TEM than in SEM. Electron diffraction and STEM showed the presence of maghemite (Fe_2O_3) and elemental iron in as-produced SWCNT. Purification procedures, previously reported in the literature, have been refined. Using TEM and STEM, spherical structures observed on the surface of SWCNT bundles were identified to be carbonaceous and probably fullerene-like. The SWCNT content, determined quantitatively using TGA and XRD, in as-produced and purified SWCNT were typically 40 and >96%, respectively.

Acknowledgments: This work has been supported by the National Science Foundation SBIR program under grant DMI-0522093. One of the authors (A. H. Reading) thanks NSF for supplemental REU funds. We are grateful to the MIT Center for Materials Science and Engineering supported by the National Science Foundation under award #DMR-0213282. Dr. Anthony Garratt-Reed from MIT is acknowledged for his assistance in STEM analysis and HyungBin Son from the Electrical Engineering and Computer Science Department at MIT for his help with Raman spectroscopy. Mr. John F. O'Brien from Nano-C is acknowledged for technical assistance.

References and Notes

1. A. Modi, N. Koratkar, E. Lass, B. Wei, and P. M. Ajayan, *Nature* 424, 171 (**2003**).
2. S. Chopra, K. McGuire, N. Gothard, A. M. Rao, and A. Pham, *Appl. Phys. Lett.* 83, 2280 (**2003**).

RESEARCH ARTICLE

3. A. Star, T.-R. Han, V. Joshi, J.-C. P. Gabriel, and G. Grüner, *Adv. Mater.* 16, 2049 **(2004)**.
4. J. N. Coleman, U. Khan, W. J. Blau, and Y. K. Gun'ko, *Carbon* 44, 1624 **(2006)**.
5. H. Koerner, G. Price, N. A. Pearce, M. Alexander, and R. A. Vaia, *Nat. Mater.* 3, 115 **(2004)**.
6. C. R. Martin and P. Kohli, *Nat. Rev. Drug Discovery* 2, 29 **(2003)**.
7. D. Pantarotto, R. Singh, D. McCarthy, M. Erhardt, J.-P. Briand, M. Prato, K. Kostarelos, and A. Bianco, *Angew. Chem. Int. Ed.* 43, 5242 **(2004)**.
8. T. Rueckes, K. Kim, E. Joselevich, G. Y. Tseng, C.-L. Cheung, and C. M. Lieber, *Science* 289, 94 **(2000)**.
9. P. Avouris, *Chem. Phys.* 281, 429 **(2002)**.
10. K. Bradley, J.-C. P. Gabriel, and G. Grüner, *Nano Lett.* 3, 1353 **(2003)**.
11. R. V. Seidel, A. P. Graham, J. Kretz, B. Rajasekharan, G. S. Duesberg, M. Liebau, E. Unger, F. Kreupl, and W. Hoenlein, *Nano Lett.* 5, 147 **(2005)**.
12. A. Kongkanand, R. Martínez Domínguez, and P. V. Kamat, *Nano Lett.* 7, 676 **(2007)**.
13. J. Giles, *Nature* 432, 791 **(2004)**.
14. S. Iijima and T. Ichihashi, *Nature* 363, 603 **(1993)**.
15. M. S. Dresselhaus, G. Dresselhaus, and P. Avouris (eds.), Carbon Nanotubes: Synthesis, Structure, Properties, and Applications, Springer, Berlin, New York **(2001)**.
16. J.-B. Donnet, R. C. Bansal, and M.-J. Wang, Carbon Black: Science and Technology, 2nd edn., Dekker, New York **(1993)**.
17. J. B. Howard, J. T. McKinnon, Y. Makarovsky, A. L. Lafleur, and M. E. Johnson, *Nature* 352, 139 **(1991)**.
18. J. B. Howard, J. T. McKinnon, M. E. Johnson, Y. Makarovsky, and A. L. Lafleur, *J. Phys. Chem.* 96, 6657 **(1992)**.
19. J. B. Howard, A. L. Lafleur, Y. Makarovsky, S. Mitra, C. J. Pope, and T. K. Yadav, *Carbon* 30, 1183 **(1992)**.
20. H. Takehara, M. Fujiwara, M. Arikawa, M. D. Diener, and J. M. Alford, *Carbon* 43, 311 **(2005)**.
21. H. Richter, J. B. Howard, and J. B. V. Sande, *Prepr. Pap.-An. Chem. Soc., Div. Fuel Chem.* 51, 91 **(2006)**.
22. W. J. Grieco, J. B. Howard, L. C. Rainey, and J. B. V. Sande, *Carbon* 38, 597 **(2000)**.
23. A. Goel, P. Hebgen, J. B. V. Sande, and J. B. Howard, *Carbon* 40, 177 **(2002)**.
24. J. B. Howard, K. D. Chowdhury, and J. B. V. Sande, *Nature* 370, 603 **(1994)**.
25. H. M. Duan and J. T. McKinnon, *J. Phys. Chem.* 98, 12815 **(1994)**.
26. H. Richter, K. Hernadi, R. Caudano, A. Fonseca, H.-N. Migeon, J. B. Nagy, S. Schneider, J. Vandooren, and P. J. V. Tiggelen, *Carbon* 34, 427 **(1996)**.
27. K. D. Chowdhury, J. B. Howard, and J. B. V. Sande, *J. Mater. Res.* 11, 341 **(1996)**.
28. M. D. Diener, N. Nichelson, and J. M. Alford, *J. Phys. Chem. B* 104, 9615 **(2000)**.
29. L. Yuan, K. Saito, W. Hu, and Z. Chen, *Chem. Phys. Lett.* 346, 23 **(2001)**.
30. L. Yuan, T. Li, and K. Saito, *Proc. Combust. Inst.* 29, 1087 **(2002)**.
31. A. V. Saveliev, W. Merchan-Merchan, and L. A. Kennedy, *Combust. Flame* 135, 27 **(2003)**.
32. R. L. Vander Wal, L. J. Hall, and G. M. Berger, *J. Phys. Chem. B* 106, 13122 **(2002)**.
33. W. Merchan-Merchan, A. V. Saveliev, and L. A. Kennedy, *Carbon* 44, 3308 **(2006)**.
34. R. L. Vander Wal and T. M. Ticich, *Chem. Phys. Lett.* 336, 24 **(2001)**.
35. R. L. Vander Wal and T. M. Ticich, *J. Phys. Chem. B* 105, 10249 **(2001)**.
36. R. L. Vander Wal, *Combust. Flame* 130, 37 **(2002)**.
37. R. L. Vander Wal, T. M. Ticich, and V. V. Curtis, *Chem. Phys. Lett.* 323, 217 **(2000)**.
38. M. J. Height, J. B. Howard, J. W. Tester, and J. B. V. Sande, *Carbon* 42, 2295 **(2004)**.
39. M. J. Height, J. B. Howard, and J. W. Tester, *Proc. Combust. Inst.* 30, 2537 **(2004)**.
40. M. J. Height, J. B. Howard, J. W. Tester, and J. B. V. Sande, *J. Phys. Chem. B* 109, 12337 **(2005)**.
41. B. S. Haynes and H. Gg. Wagner, *Prog. Energy Combust. Sci.* 7, 229 **(1981)**.
42. J. Z. Wen, H. Richter, W. H. Green, J. B. Howard, M. Treska, P. M. Jardim, and J. B. Vander Sande, *J. Mater. Chem.* 18, 1561 **(2008)**.
43. K. Tohji, H. Takahashi, Y. Shinoda, N. Shimizu, B. Jeyadevan, I. Matsuoka, Y. Saito, A. Kasuya, S. Ito, and Y. Nishina, *J. Phys. Chem. B* 101, 1974 **(1997)**.
44. A. G. Rinzler, J. Liu, H. Dai, P. Nikolaev, C. B. Huffman, F. J. Rodríguez-Macías, P. J. Boul, A. H. Lu, D. Heymann, D. T. Colbert, R. S. Lee, J. E. Fischer, A. M. Rao, P. C. Eklund, and R. E. Smalley, *Appl. Phys. A* 67, 29 **(1998)**.
45. I. W. Chiang, B. E. Brinson, R. E. Smalley, J. L. Margrave, and R. H. Hauge, *J. Phys. Chem. B* 105, 1157 **(2001)**.
46. Nano-C, Inc., www.nano-c.com.
47. M. S. Dresselhaus, G. Dresselhaus, A. Jorio, A. G. S. Filho, and R. Saito, *Carbon* 40, 2043 **(2002)**.
48. A. Jorio, M. A. Pimenta, A. G. S. Filho, R. Saito, G. Dresselhaus, and M. S. Dresselhaus, *New Journal of Physics* 5, 139.1 **(2003)**.
49. M. S. Dresselhaus, G. Dresselhaus, R. Saito, and A. Jorio, *Physics Reports* 409, 47 **(2005)**.
50. A. C. Ferrari and J. Robertson, *Phys. Rev. B* 61, 14095 **(2000)**.
51. M. Milnera, J. Kürti, M. Hulman, and H. Kuzmany, *Phys. Rev. Lett.* 84, 1324 **(2000)**.
52. R. Martel, H. R. Shea, and P. Avouris, *J. Phys. Chem. B* 103, 7551 **(1999)**.
53. A. E. Cohen and L. Mahadevan, *Proc. Natl. Acad. Sci., USA* 100, 12141 **(2003)**.
54. K. Sai Krishna and M. Eswaramoorthy, *Chem. Phys. Lett.* 433, 327 **(2007)**.
55. D. Hecht, L. Hu, and G. Grüner, *Appl. Phys. Lett.* 89, 133112 **(2006)**.
56. M. J. O'Connell, S. M. Bachilo, C. B. Huffman, V. C. Moore, M. S. Strano, E. H. Haroz, K. L. Rialon, P. J. Boul, W. H. Noon, C. Kittrell, J. Ma, R. H. Hauge, R. B. Weisman, and R. E. Smalley, *Science* 297, 593 **(2002)**.
57. M. F. Islam, E. Rojas, D. M. Bergey, A. T. Johnson, and A. G. Yodh, *Nano Lett.* 3, 269 **(2003)**.
58. T. Hertel, A. Hagen, V. Talalaev, K. Arnold, F. Hennrich, M. Kappes, S. Rosenthal, J. McBride, H. Ulbricht, and E. Flahaut, *Nano Lett.* 5, 511 **(2005)**.

Received: 1 June 2007. Revised/Accepted: 5 October 2007.

Transplanting assembly of carbon-nanotube-tipped atomic force microscope probes

Soohyung Kim, Hyung Woo Lee, and Sang-Gook Kim[a)]
Department of Mechanical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, Massachusetts 02139, USA

(Received 27 December 2008; accepted 22 April 2009; published online 13 May 2009)

Carbon-nanotube (CNT)-tipped atomic force microscope (AFM) probes were assembled in a deterministic and reproducible manner by transplanting a CNT bearing polymeric carrier to a microelectromechanical systems cantilever. Single-strand CNTs were grown vertically at predefined locations where each CNT was encapsulated into a cylindrical polymer carrier block. Double-layer carriers were used for controlling the release of blocks and the exposed length of CNT tips after the assembly. Much reduced complexity in assembly was achieved by transplanting individual CNTs to AFM probes, which could scan nanotrenches and biostructures with little probe artifacts. © *2009 American Institute of Physics*. [DOI: 10.1063/1.3136762]

Carbon nanotubes (CNTs) have attracted much attention due to their excellent mechanical, electrical, chemical, and thermal properties. Various potential applications have been suggested and demonstrated,[1] which would require the assembly of CNTs with a control of their location, orientation, and geometrical configurations. Promising efforts have been reported to locate and assemble nanostructures such as CNTs with the Langmuir–Blodgett technique, fluid flow and capillary forces, external forces, and templates, among many.[2–4] An assembly of some functional nanodevices, however, requires locating an individual nanostructure at a deterministic position of micro- or mesoscale structures, which is still a very complex task. A CNT-tipped probe for an atomic force microscope (AFM) is a typical application, which requires an assembly of a single-strand CNT at the tip of an AFM cantilever. The degree of complexity increases rapidly as the assembly system's scale order grows, which can be defined as the logarithmic ratio of the size of the target assembly system to the smallest characteristic length of its nanoscale component. In this letter, we report that transplanting assembly of individual CNTs can reduce the complexity of assembly by encapsulating individual CNTs into microcarrier blocks and demonstrate that this method can be used to make CNT-tipped AFM probes efficiently.

The nanometer-scale diameter, high-aspect-ratio cylindrical geometry, easy buckling under excessive load, and superior wear resistance of CNTs make the CNT-tipped AFM probe a tool for high-resolution imaging and scanning of nano- and biostructures.[5,6] However, the volume manufacture of these has been a challenging task due to the complexity of handling and locating only one CNT tip at the apex of an AFM cantilever. Previous approaches to the fabrication of CNT-tipped AFM probes include attaching a CNT at the apex of an AFM probe manually,[5] with a guidance of the external magnetic[7] and electric fields,[8] pick-and-place using an AFM device,[9] or growing CNTs at embedded catalytic metal seeds at the apex of AFM probes.[10] All approaches showed good functionality of CNT-tipped AFM probes but required time-consuming effort such as removing redundant CNT tips and shortening, aligning, or welding CNT tips.

Transplanting a bundle of CNTs was demonstrated by one of the authors previously.[11] By decoupling growing from assembling CNTs, transplanting assembly showed a potential that a high rate production of CNT-based devices could be achieved. But the previous method could not handle or assemble individual CNTs. We report a method for the deterministic assembly of individual CNTs, which consists of growing an array of vertically aligned single-strand CNTs, encapsulating each CNT into a microscale polymer carrier block, which can be handled and manipulated with existing microscale tools, and transplanting the CNT-bearing polymer blocks to target locations followed by the exposure of buried CNTs. Figure 1 summarizes the transplanting assembly process we designed to make a CNT-tipped AFM probe. The geometry and the orientation of individual CNTs are frozen into photosensitive polymer layers spin casted over an array of single-strand CNTs vertically grown at the predefined locations on a silicon substrate. Then the photosensitive polymer layers are lithographically patterned to form cylindrical blocks over the individual CNTs, which is similar to the cookie-cutting process. An intentional undercut of the bottom layer holds the carrier blocks until the release from the substrate and controls the exposed length of the CNT tips after assembly.

The first step in transplanting assembly is the vertical growth of CNTs, which requires seeding the nickel (Ni) catalytic nanodots at predefined locations on the Si substrate. A 21×21 array of Ni catalytic dots (100–200 nm in diameter and 30 nm in thickness) was defined using electron beam lithography (Raith 150 at MIT's Scanning Electron Beam

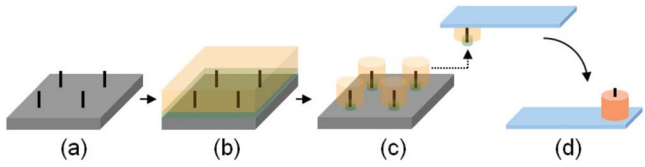


FIG. 1. (Color online) Transplanting assembly process steps for CNT-tipped AFM probes. (a) Vertical CNT growth, (b) encapsulation, (c) cookie cutting (forming individual polymer block containing a single CNT tip at the bottom with intentional undercut), and (d) CNT tip release.

[a)]Author to whom correspondence should be addressed. Electronic mail: sangkim@mit.edu.

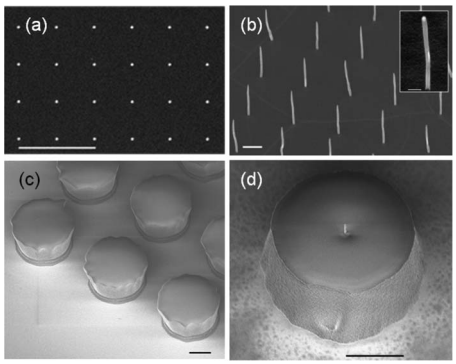


FIG. 2. The growth and encapsulation of an array of vertically aligned individual CNTs. (a) An array of 30 nm thick Ni dots formed by electron beam lithography (scale bar: 10 μm). (b) An array of vertically grown CNT strands (scale bar: 2 μm). The inset shows a freestanding CNT, 150 nm in diameter and 5 μm in length (scale bar: 500 nm). (c) An array of SU8 blocks encapsulating CNTs (scale bar: 10 μm). (d) An SU8 block shows a single CNT after release (scale bar: 10 μm).

Lithography Facility) followed by metal deposition and lift-off process. An array of vertically aligned CNTs was grown using a home-built plasma enhanced chemical vapor deposition machine.[12] Each CNT strand was then embedded into a micropolymer block, which serves as a CNT carrier. We used a double polymeric layer encapsulation process with SU8 (top: MicroChem) and polymethylglutarimide (PMGI) (bottom: MicroChem): the top SU8 (15 μm in thickness and 20 μm in diameter) forms the body of the carrier, while the bottom PMGI layer (1.5 μm in thickness) holds the body until the release of the carrier from the substrate and then is removed to expose the CNT tip. Figure 2 shows Ni catalytic dots defined using electron beam lithography [Fig. 2(a)], an array of vertically grown CNT strands [Fig. 2(b)], an array of SU8 pillars encapsulating individual CNTs [Fig. 2(c)], and an inverted SU8 block bearing a single CNT tip [Fig. 2(d)]. The orientation of the embedded CNT is near parallel to the axis of the SU8 block. The diameter of CNTs matches the size of Ni dots, and the length is 5–10 μm with a uniform cylindrical shape. The thickness of the bottom PMGI layer was chosen to be 1.5 μm so that a target aspect ratio of the CNT tip is about 10, which would give about 0.1 nm thermal fluctuation at the end of the tip at room temperature (300 K). The 20 μm diameter of SU8 blocks was the result of process constraints in our facility, which can be further reduced to achieve the higher resonance frequency and the higher aspect ratio of the probes.

We investigated the physicochemical interaction of individual CNTs to see whether the pristine properties of CNTs can be preserved during the transplanting process, and the vertical orientation of grown CNTs can be maintained under the fluidic shearing during the spin coating of polymers. High-resolution transmission electron microscopy pictures show the same graphene layers with 0.34 nm spacing before and after the encapsulation process, indicating that pristine graphene structures are intact throughout the encapsulation processes. In order to predict the flow-induced deflection during the spin coating of polymers, the polymer spin-coating process is modeled as a one-dimensional (radial direction) laminar flow by the centrifugal force and shearing force, as shown in

$$U(r,z) = \frac{\omega^2 R z}{\nu}\left(t - \frac{z}{2}\right), \tag{1}$$

where ω is the rotational speed, R is the distance from the center of rotation, z is the height from the substrate, ν is the viscosity of a polymer, and t is the thickness of the polymer layer. The corresponding Reynolds number is about 5×10^{-4} at the maximum velocity, indicating a laminar flow. The array of vertically aligned CNTs packed within a 1×1 mm^2 area is located near the central position of the spin axis, and each CNT is modeled as a rod rigidly clamped to the substrate with the assumption of no-slip condition on the substrate. Drag coefficients are calculated using the radial velocity distributions along the CNT, and force distributions are obtained from the drag coefficients in

$$C_D = \left(\frac{8\pi}{\mathrm{Re}\,\ln(7.4/\mathrm{Re})}\right)\left(\frac{3 + 2\phi^{5/3}}{3 - 4.5\phi^{1/3} + 4.5\phi^{5/3} - 3\phi^2}\right), \tag{2}$$

$$F_D = \tfrac{1}{2} C_D \rho U^2 A, \tag{3}$$

where ϕ is the areal fill factor of CNT strands in each cell. The final deflection at the end of a CNT tip is calculated using a linear beam bending model under distributed loads, which predicts less than 1° deflection from the vertical axis. This small deflection with respect to the dimension of CNT tips (150 nm in diameter and 1.5 μm in length) is negligible and matches with the observation at the transplanted CNT tips [Fig. 2(d)].

After encapsulation, a polymer block where a single-strand CNT is embedded is transplanted manually to the end of a silicon cantilever using the micromotion probe stage under an optical microscope. A tipless AFM cantilever (NCS12, MicroMasch), which already picked a drop of liquid adhesive (LOCTITE® 408), approaches a target pellet. When the adhesive is dried, the pellet attached to the end of the AFM cantilever is released by gently shearing the pellet. After assembly, we etch any remaining PMGI bottom layer and expose a CNT tip of about 1.5 μm, which corresponds to the bottom (PMGI) layer thickness. Various CNT-tipped AFM probes for different scanning operation modes were fabricated by assembling a CNT embedded SU8 pellet on various AFM cantilevers. Figure 3 shows AFM probes with a transplanted single-strand CNT. The fact that even a manual assembly of CNT is feasible shows that the complexity of assembly has been reduced by encapsulating CNTs into microcarrier blocks. CNT-tipped AFM probes for different modes of scanning could be produced by transplanting assembly of individual CNTs on various cantilevers (Si- and Au-coated Si_3N_4 cantilevers), which demonstrates the flexibility of transplanting assembly as a viable manufacturing process for CNT-tipped probes.

The performance of transplanting-assembled AFM probes was tested by mounting them on a commercial AFM (DI 3100, Veeco Instruments Inc.) and scanning over an AFM calibration grating (TGZ02, MikroMasch). The Si grating consists of 106 nm deep vertical trenches with a period of 3 μm. We used the contact mode scanning with a scanning range of 10×10 μm^2 and a scanning speed of 10 μm/s. The transplanting-assembled CNT AFM probes scanned the vertical walls of trenches with sidewall angles larger than 85°, which shows much less probe artifact than

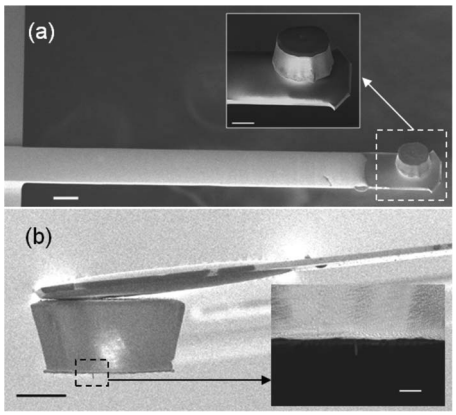


FIG. 3. CNT-tipped AFM probes made by transplanting assembly. (a) A contact mode probe on a Si cantilever with a spring constant of 0.3 N/m (scale bar: 20 μm). The inset shows the enlarged view of the CNT tip (scale bar: 10 μm). (b) The side view of a tapping mode CNT-tipped AFM probe with a tuned resonance frequency of 60 kHz (scale bar: 10 μm). The inset shows the vertical CNT tip (scale bar: 2 μm).

that with the standard Si probe (Fig. 4). We also scanned soft protein structures such as actin filament (F-actin) samples, as shown in Fig. 5. F-actin was prepared in a buffer solution and stored at 4 °C before scanning. 10 ml of the solution was dropped on a glass slide, and AFM scanning was performed after the solution had dried.

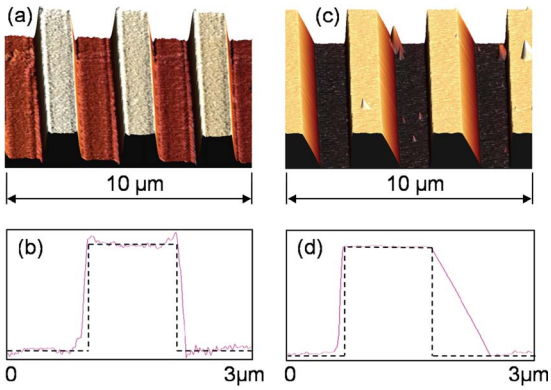


FIG. 4. (Color online) Comparison of AFM scanning results on an AFM calibration grid. (a) A scanning result with a CNT-tipped AFM probe (contact mode). (b) The sectional profile of (a) where the dotted line represents the ideal shape of 106 nm deep trenches and the solid line denotes the scanning results. (c) Scanning result with a standard Si AFM probe (tapping mode). (d) The sectional profile of (c).

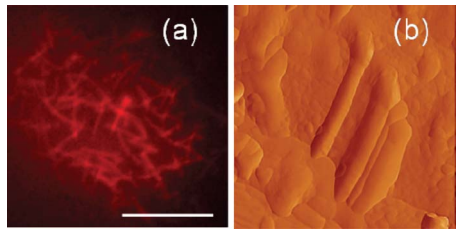


FIG. 5. (Color online) The scanning of a biological sample (64 μl/2.3 μM of F-actin) with a CNT-tipped AFM probe. (a) Fluorescent microscopy shows bundles of F-actins in the buffer solution (scale bar: 10 μm). (b) AFM scanning image of F-actins (scanning range: 3×3 μm^2).

We demonstrated that transplanting assembly could assemble individual CNTs into microscale cantilevers at much reduced complexity. This could be achieved by shifting the scale order of assembly from "nano/micro" to "micro/micro," thus reducing the complexity of the assembly. Transplanting assembly enables even manual assembly of a CNT-tipped AFM probe to be done in minutes excluding the adhesive curing time. We believe that this assembly method can be scaled up and automated to make a massive parallel assembly of nanostructures for high throughput applications.

The authors thank Intelligent Microsystems Center in Korea, DARPA Grant No. HR0011-06-1-0045, and Hewlett Packard for a partial funding of this research. The first and the second author contributed equally to this work.

[1]R. H. Baughman, A. A. Zakhidov, and W. A. de Heer, Science **297**, 787 (2002).
[2]Y. Huang, X. Duan, Q. Wei, and C. M. Lieber, Science **291**, 630 (2001).
[3]Y. Cui, M. T. Bjork, J. A. Liddle, C. Sonnichsen, B. Boussert, and A. P. Alivisatos, Nano Lett. **4**, 1093 (2004).
[4]P. A. Smith, C. D. Nordquist, T. N. Jackson, T. S. Mayer, B. R. Martin, J. Mbindyo, and T. E. Mallouk, Appl. Phys. Lett. **77**, 1399 (2000).
[5]H. Dai, J. H. Hafner, A. G. Rinzler, D. T. Colbert, and R. E. Smalley, Nature (London) **384**, 147 (1996).
[6]T. Larsen, K. Moloni, F. Flack, M. A. Eriksson, M. G. Lagally, and C. T. Black, Appl. Phys. Lett. **80**, 1996 (2002).
[7]A. Hall, W. G. Matthews, R. Superfine, M. R. Falvo, and S. Washburn, Appl. Phys. Lett. **82**, 2506 (2003).
[8]J. Tang, G. Yang, Q. Zhang, A. Parhat, B. Maynor, J. Liu, L. Qin, and O. Zhou, Nano Lett. **5**, 11 (2005).
[9]J. H. Hafner, C. Cheung, T. H. Oosterkamp, and C. M. Lieber, J. Phys. Chem. B **105**, 743 (2001).
[10]Q. Ye, A. M. Cassell, H. Liu, K. Chao, J. Han, and M. Meyyappan, Nano Lett. **4**, 1301 (2004).
[11]T. El-Aguizy, J.-h. Jeong, Y. B. Jeon, W. Z. Li, Z. F. Ren, and S. G. Kim, Appl. Phys. Lett. **85**, 5995 (2004).
[12]Z. F. Ren, Z. P. Huang, J. W. Xu, J. H. Wang, P. M. Bush, P. Siegal, and P. N. Provencio, Science **282**, 1105 (1998).

Journal of The Electrochemical Society, **154** (10) H887-H893 (2007)
0013-4651/2007/154(10)/H887/7/$20.00 © The Electrochemical Society

Memory Functions of Nanocrystalline Indium Tin Oxide Embedded Zirconium-Doped Hafnium Oxide MOS Capacitors

Adam Birge,* Chen-Han Lin,* and Yue Kuo,z**

Thin Film Nano and Micro Research Laboratory, Department of Chemical Engineering, Texas A&M University, College Station, Texas 77843-3122, USA

A floating gate metal-oxide-semiconductor capacitor memory device utilizing nanocrystalline indium tin oxide (ITO) layer embedded in a zirconium-doped hafnium oxide (Zr-doped HfO_2) high-k gate dielectric has been fabricated and studied. The embedded ITO layer has crystalline structure with a grain size of about 5.4 nm. Capacitance-voltage and current-voltage measurements show positive charge trapping under the negative gate bias operation condition. Comparison to a control sample shows a fourfold increase (5.3×10^{-12} cm^{-2} to 1.4×10^{-12} cm^{-2}) in oxide trapped charge density and opposite trapped charge polarity, indicating that the observed effects are due to the inclusion of the ITO floating gate layer. The device maintains a large postwrite window (>100 mV) over a ten-year period. Furthermore, the prominent charge transfer mechanism is direct tunneling. The negative differential resistance in the current-voltage curve shows the existence of the coulomb blockade effect that may limit negative charge storing and retention. The asymmetrical barrier of the Zr-doped HfO_2 allows for the enhanced hole retention while eliminating the possibility of electron retention.
 [DOI: 10.1149/1.2768291]

Manuscript submitted April 26, 2007; revised manuscript received June 13, 2007. Available electronically August 15, 2007.

Traditionally, the flash memory device has a metal-oxide-semiconductor (MOS) field-effect transistor structure where a layer of polysilicon is embedded into the gate oxide to serve as a charge trapping medium.[1-3] The dielectric layer between the control gate and floating polysilicon gate, known as the control oxide, needs to be thick enough to prevent charge transfer between these two gates under the strong control gate bias condition. At the same time, the dielectric layer between the floating gate and substrate, known as the tunneling oxide, should to be thin enough to allow easy transfer of charges when a strong gate bias voltage is applied. Under the normal gate bias condition, charges cannot tunnel back from the floating gate to the substrate due to the existence of a potential barrier between the floating gate and the substrate. However, in the retention mode, the device should be able to store charges for ten years or more. The existence of a charge trapping layer embedded in the gate dielectric structure creates a discontinuity in the electric field, which alters the flatband voltage (V_{FB}) and the threshold voltage (V_T) of the device.[4,5] Thus, by transferring electrons or holes to the floating gate the memory cell can be alternated between "1" and "0" logical states, and these states can be observed by measuring the change of the drain current or the threshold voltage shift (ΔV_T) of the transistor. The gate bias for reading is chosen so that the transistor will be on while in either the 1 or 0 memory state and off while in the alternate state.

As device dimensions scale down to allow for a greater circuit density, a higher switching speed, and a lower operating power, the conventional flash structure begins to face severe problems.[1,2,6] Due to the large electron barrier of silicon dioxide (SiO_2), high voltage is required for high-speed write and erase processes. Reducing the voltage will degrade the writing efficiency and therefore the write time. When the SiO_2 thickness is reduced to the nanometer range, any point defects can create a leakage channel from the floating gate to substrate. In addition, for this kind of ultrathin oxide layer, defects or percolation paths can be easily generated from the stress-induced leakage current mechanism.[7] Since the conventional floating gate is a continuous conducting layer, a single leakage path through the tunneling oxide can quickly drain charges stored on the entire floating gate site. In order to eliminate the issue of a continuous conducting sheet discharging through a single leakage path, several alternate methods have been proposed to modify the polysilicon floating gate structure.[8,9] Recently, one of the most widely researched structures is the nanocrystal embedded floating gate structure.[8,10,11] For this kind of device, the floating gate is divided into many discreet nanocrystalline nodes. The nodes are isolated from each other by the surrounding dielectric material so that a single leakage path will only disturb one or a few nodes localized to the path but not the rest. Although the earliest works were concentrated on using silicon nanocrystals as the charge-storage media,[10] germanium[12] or germanium–silicon[13] nanocrystals are also frequently used as the small germanium bandgap can create a deeper storage well. Although charges can be stored at the conduction or valence band edges or at deep traps of the bandgap, the semiconductor nanocrystals suffer from low density of states and relatively shallow quantum wells. Recently metal nanocrystals have been studied as an alternative charge storage media[14,15] because they have large density of states, various work functions (and therefore well depths), and enhanced tunnel oxide electric fields in the immediate vicinity of the nanocrystal.[16]

A conducting metal oxide nanocrystal may provide benefits of both the semiconductor and metal nanocrystals. For example, indium tin oxide (ITO) is a degenerately doped n-type semiconductor with a large work function and large bandgap.[17,18] When ITO is embedded in a gate dielectric layer, the large density of states near the Fermi level and high work function may offer similar characteristics as those of embedded metals, while the presence of a large bandgap could allow for sub-Fermi level traps for greater charge retention.

Many researchers have been investigating replacement of the SiO_2 insulating layer in high performance metal-oxide-semiconductor (MOS) devices with a high dielectric constant (high-k) material because a physically thicker layer can be used while maintaining similar electrical properties.[19] Additionally, high-k dielectrics generally have bandgaps that decrease with increasing k value.[20] With a smaller bandgap, the barriers for electron and hole tunneling would be smaller, allowing for lower operating voltages and speed in write and erase operations.[21] Recently, it has been demonstrated that doping the high-k material with a proper amount of a third element can improve many of its dielectric properties.[22,23] For example, the doped high-k film can have a higher amorphous-to-polycrystalline temperature and a lower interface layer thickness than the undoped high-k film.[24,25] The polycrystalline dielectric film contains grain boundaries that enhance leakage current and can be a potential reliability problem. A thick interface layer can cause a large equivalent oxide thickness (EOT) in the high-k stack. Among all conventional high-k materials, currently hafnium oxide (HfO_2) and zirconium oxide (ZrO_2) are two of the most studied materials. The Zr-doped HfO_2 high-k film has many improved dielectric properties over those of HfO_2 or ZrO_2.[26,27] In this study, the authors

* Electrochemical Society Student Member.
** Electrochemical Society Fellow
z E-mail: yuekuo@tamu.edu

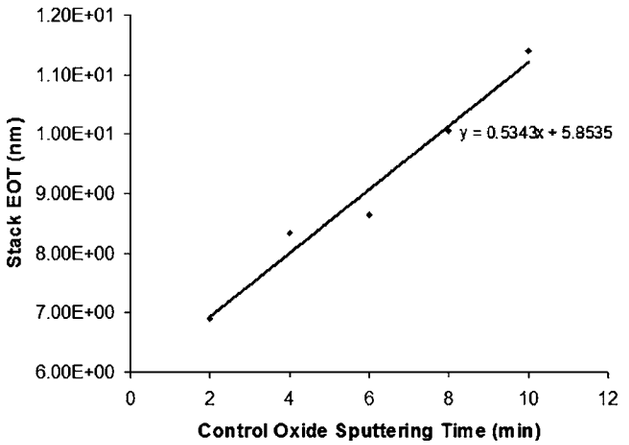


Figure 1. Stack EOT for 2 min tunnel oxide, 1 min ITO floating gate layer, and variable control oxide sputtering time.

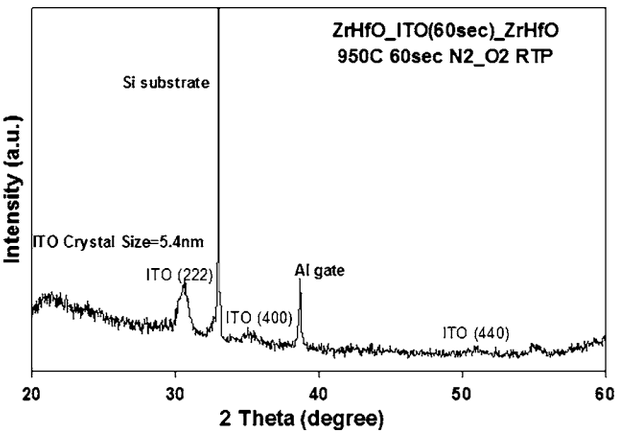


Figure 2. XRD of ITO embedded Zr-doped HfO_2 sample.

fabricated MOS capacitors using the Zr-doped HfO_2 as the gate dielectric layer with an embedded nanometer-thick ITO layer and investigated the process influence on the memory functions.

Experimental

The Zr-doped HfO_2 (tunneling oxide)/ITO/Zr-doped HfO_2 (control oxide) trilayer structure was sputter deposited on a p-type (10^{15} cm^{-3}) silicon wafer in one pump down. The Zr-doped HfO_2 layers were deposited from a composite Zr–Hf target at 5 mTorr and Ar/O2 1:1 ratio. The ITO layer was deposited from an ITO target at 5 mTorr and Ar gas. The bottom tunneling oxide and ITO deposition times were fixed at 2 min and 1 min, respectively. The control oxide deposition time was varied at 2 min (S2), 4 min (S4), 6 min (S6), 8 min (S8), or 10 min (S10), respectively. The entire embedded high-*k* stack was annealed for 1 min at 950°C under N_2/O_2 (1:1) ambient by rapid thermal annealing (MTP RTP 600S). The capacitor gate electrode was composed of 200 nm thick sputtered aluminum (Al). The backside native oxide of the wafer was stripped with a HF solution and then deposited with Al for better contact. The complete capacitor was annealed at 300°C for 5 min under H_2/N_2 atmosphere. For the comparison purpose, a control capacitor that contains no ITO embedded layer in the Zr-doped HfO_2 gate dielectric structure was prepared.

The capacitor's capacitance-voltage (*C-V*) and current-voltage (*I-V*) characteristics were measured with a HP4284A *LCR* meter and HP4155C semiconductor parameter analyzer, respectively. The NCSU CVC program[28] was used to estimate the capacitor's EOT and V_{FB} values. The top control oxide's EOT was extracted from a linear fit of the sputtering time vs the complete high-*k* stack's EOT as shown in Fig. 1. The bottom tunneling oxide's EOT was approximately 1.96 nm on all samples. The additional 3.9 nm EOT difference is a combination of the EOT of the floating gate layer, which is a mixture of ITO and dielectric, and the interfacial layer formation at the Si/high-*k* interface. The samples were analyzed with X-ray diffraction (XRD) for the crystallinity of the embedded ITO layer. Figure 2 shows that the embedded ITO is crystalline with predominate (222) orientation. The crystal size is about 5.4 nm, determined from the peak location and full width at half maximum.[29]

Results and Discussion

Charge trapping characteristics.— *C-V* results from the samples are summarized in Fig. 3. To determine which sample displayed the best memory characteristics, the *C-V* curves were measured from the accumulation (negative gate bias voltage) region to the inversion (positive gate bias voltage) region and back to the accumulation region for each sample on the ranges of (−2 V, 2 V, − 2 V), (−3 V, 3 V, − 3 V), (−5 V, 5 V, − 5 V), and (−7 V, 7 V, − 7 V). The hysteresis was measured as the change in V_{FB} (ΔV_{FB}) from the forward sweep (accumulation to inversion) to the return sweep (inversion to accumulation). Also, the hysteresis of the control sample (about −100 mV at a 7 V magnitude sweep range) is shown for comparison. For the desired case of charge being injected from the substrate to the floating gate, a positive value of hysteresis is expected for the p-type substrate, as only positive charge can be injected in accumulation and only negative charge can be injected in inversion.

From the data in Fig. 3, it is obvious that the charge injection and trapping mechanisms are strongly dependent on the physical thicknesses of the dielectric layers. The hysteresis for the samples with the thinner control oxides, S2 (EOT 1.07 nm) and S4 (EOT 2.14 nm), is in the positive direction. The lesser magnitude of hysteresis for sample S2 as compared to S4 is most likely due to the extremely thin control oxide layer of 1.07 nm EOT. At this thickness, the direct tunneling mechanism across the control oxide will inhibit the amount of charge that can be stored in the ITO site due to electrons tunneling from the Al gate to the ITO site and due to excess holes tunneling from the ITO site to the Al gate. The thicker control oxide samples, S8 (EOT 4.27 nm) and S10 (EOT 5.34 nm), have hysteresis trends in the opposite direction from those observed in S2 and S4, and the magnitude of the hysteresis increases with the sweep range. S6, with control oxide EOT of approximately 3.21 nm,

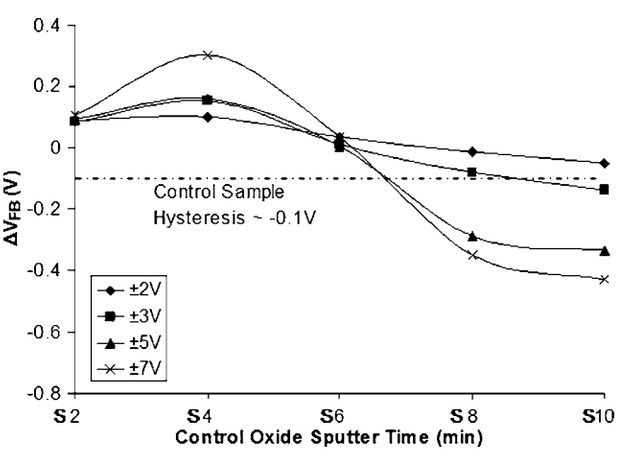


Figure 3. Hysteresis of samples prepared at different control oxide sputtering times and swept at different voltage ranges. The control sample's hysteresis swept at the −7 to +7 V range is approximately −0.1 V.

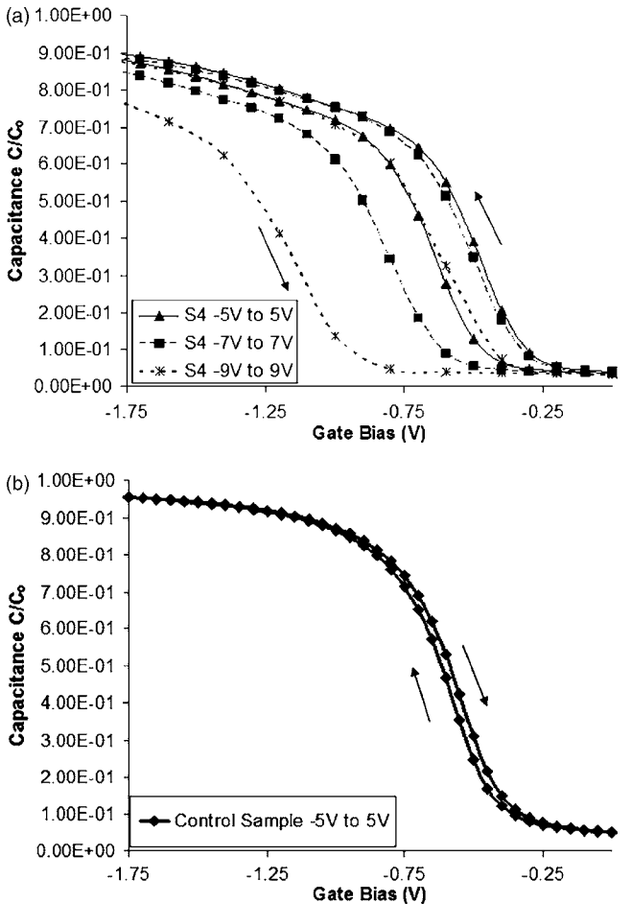


Figure 4. (a) *C-V* curves for sample S4 at various sweep ranges and (b) control sample hysteresis at −5 to 5 V sweep range.

has almost no hysteresis regardless of the sweep range. A general trend can be seen for S4 through S10 where increasing control oxide thickness results in the hysteresis becoming progressively more negative. The thicker control oxides account for an increasingly larger portion of the voltage drop across the entire dielectric, so that the electric field across the tunnel oxide for a given gate bias will be weaker for samples with thicker control oxides. The control sample exhibited a negative hysteresis indicating inherent bulk traps; therefore the ITO-embedded samples should also exhibit negative hysteresis if the applied bias is not strong enough for significant hole trapping. It is also possible that the density and distribution of the inherent bulk charges could play a role in the observed trend. Since sample S4 (control oxide EOT ~2.14 nm) has shown the most desirable hysteresis characteristic, it has been chosen for more detailed study in the rest of this paper.

C-V hysteresis for sample S4 are shown in Fig. 4a for different voltage sweep ranges. A large counterclockwise hysteresis in the *C-V* curve can be seen, and the magnitude of this hysteresis increases with the sweep voltage range, e.g., 161, 305, and 780 mV for (−5 V, 5 V, − 5 V), (−7 V, 7 V, − 7 V), and (−9 V, 9 V, − 9 V) sweeps, respectively. It is also seen that for sweeps that start at stronger negative bias, both the forward and return sweep shifted further left. To determine if the hysteresis is due to the presence of the embedded ITO nanolayer, the *C-V* curve of the control sample (no ITO) was measured. Figure 4b shows the *C-V* characteristics of the control sample in the same scale, which shows 41 mV hysteresis for a (−5 V, 5 V, − 5 V) sweep. The density of oxide trapped charge (Q_{OT}) can be approximated by the product of capacitance (33 pF for S4) and ΔV_{FB}.[4,5] Therefore, the density of trapped charges in S4, the density of trapped charge for the ITO-embedded sample S4, 5.3×10^{-12} eV per cm^2, is nearly four times as large for a (−5 V, 5 V, − 5 V) sweep than for the control sample, which has an approximate 1.4×10^{-12} eV per cm^2 charge density. In fact, the hysteresis of the control sample, while very small, is even in the opposite direction from the ITO-embedded sample. This indicates that the charge trapping mechanism seen in Fig. 4a is due to the existence of the embedded ITO layer and not an inherent feature of the Zr-doped HfO_2 high-*k* dielectric. Holes are trapped in the bulk ITO nanolayer or at the ITO/high-*k* interface.

For counterclockwise hysteresis in a MOS capacitor, a net positive charge must be introduced into the oxide at negative bias, or a net negative charge must be introduced into the oxide at positive bias. The charge transfer can take place to the ITO site from the gate or substrate, or conversely can be the result of charges transporting from the ITO site to the gate or substrate. As a qualitative look at the source and type of injected charges, *C-V* curve hysteresis can be measured from different swept voltage ranges, for example, in the low bias range, the low positive bias-to-high negative bias range, and the low negative bias-to-high positive bias range, and compared.[30] Figures 5a-c show these results on sample S4. For very low bias sweep ranges (−1.25 to 0.25 V) the hysteresis is very small and in the counterclockwise direction, as shown in Fig. 5a. A sweep from small positive bias to large negative bias (0.25 to − 6 V) results in a much larger hysteresis than the previous case but in the same counterclockwise direction, as shown in Fig. 5b. However, from small negative bias to large positive bias (−1 to 6 V), the hysteresis is small, as shown in Fig. 5c, like the low bias range case. The fact that the counterclockwise hysteresis is enhanced predominately by negative bias indicates that the memory effect is due to a net positive charge trapping. In addition, the backward sweep directions all have a negative V_{FB} shift from the fresh case, which indicates that the charge erase efficiency at the positive bias is low. The band diagram of the MOS memory capacitor showing the write and erase conditions are summarized in Fig. 6a and b, respectively. Additionally, Fig. 6c represents the band diagram of the capacitor under low bias retention conditions, where the net positive charge is trapped in the ITO nanolayer.

Trapped charge retention.— The ability to create a controlled *C-V* hysteresis, and therefore, an adjustable threshold voltage, can only be exploited for nonvolatile memories if the injected charge can maintain long-term retention. Retention was measured by periodically performing a small-signal *C-V* measurement after a write or erase pulse. The V_{FB} can then be extracted from these curves and compared to the fresh value to determine the amount of voltage shift.

Figure 7 displays the retention data for an S4 sample after various write and erase pulses. As illustrated in Fig. 7a, the pulses were performed in the following order: (1) 5 s erase at +7 V, (2) 5 s write at −7 V, (3) additional 5 s write at −7 V, and (4) final 10 s erase at +7 V. After each pulse the retention was measured over a time period of 5000 s. It is clear from Fig. 7b that an erase (positive pulse) on a fresh sample, i.e., curve 1, results in very little charge storage, which is the same effect seen in *C-V* sweeps to strong positive bias. However, a subsequent fresh write cycle results in a large initial flatband voltage shift of about 230 mV, i.e., curve 2. Using the approximation of a two-dimensional oxide sheet charge located away from the interface,[4] the number of initial trapped charges can be calculated as

$$N_t = \gamma \times (C_{ox}/q) \times (-\Delta V_{FB}) \quad [1]$$

with C_{ox} equal to the capacitance at accumulation, q is the electron charge, and a resultant flatband voltage shift of ΔV_{FB}; γ is a correction factor to take in account the distance of the charge from the substrate interface. If the ITO nanocrystal diameter and density are known, Eq. 1 can take on a more complex form for more accurate correction.[10] An estimation can be made by setting γ equal to the

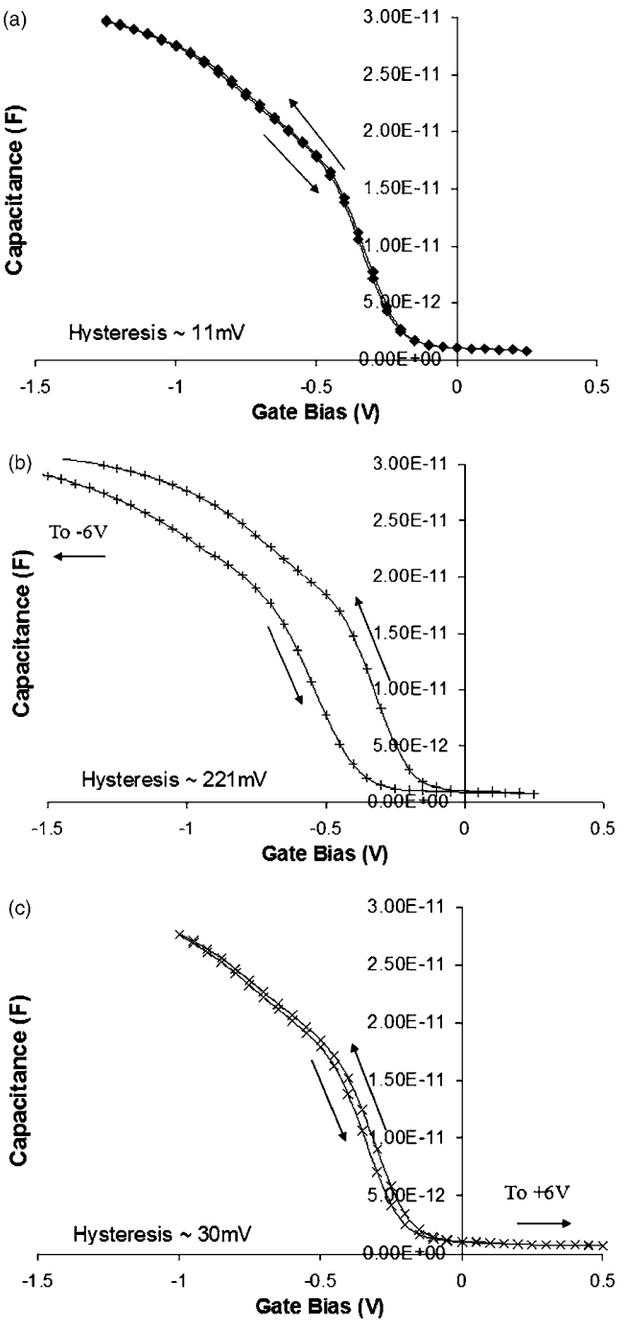


Figure 5. *C-V* curves of sample S4. (a) small swept voltage range (−1.25 to 0.25 V), to (b) large negative bias range (0.25 to −6 V), and to (c) large positive bias range (−1 to 6 V).

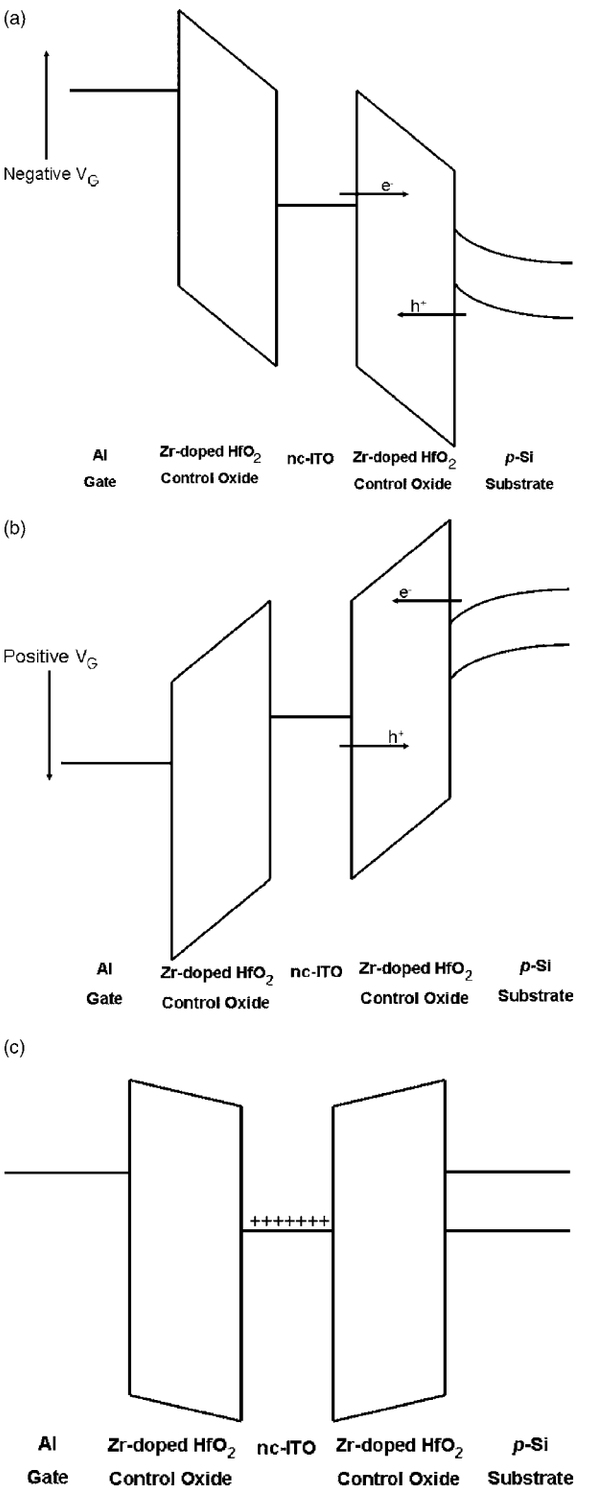


Figure 6. Band diagram under (a) negative bias writing conditions, (b) positive bias erasing conditions, and (c) small bias retention conditions with positive trapped charge.

ratio of the sum of control and tunnel oxide EOTs to the control oxide EOT. For example, using C_{ox} = 33 pF, $q = 1.6 \times 10^{-19}$ C, initial $\Delta V_{FB} = -230$ mV, and γ as 1.92, N_t is calculated to be $\sim 9.12 \times 10^7$. With the capacitor diameter at 100 μm, this corresponds to a two-dimensional trapped hole density of 1.16×10^{12} cm^{-2}. Spherical shaped nanocrystals have a maximum two-dimensional packing density of 90.69%, so that 3.96×10^{12} nanocrystals per cm^2 would be present in a maximum density configuration of 5.4 nm diam nanocrystals. Assuming a nonmaximal nanocrystal distribution, the 1.16×10^{12} cm^{-2} calculated trapped charge density is reasonable for one charge per nanocrystal.

Curve 2 displays excellent retention characteristics. The retention data can be closely fitted by a logarithmic approximation that yields a decay rate of 17 mV/decade, yielding a flatband voltage that is still 109 mV away from the fresh value after a time period of 10^8 s (~10 years). A second write cycle, i.e., curve 3, shows even greater

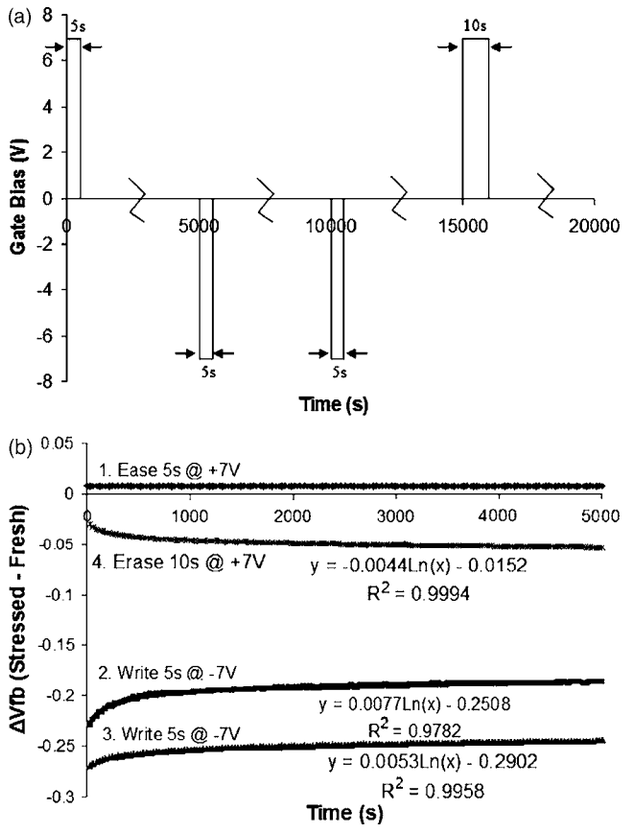


Figure 7. (a) Pulse sequence for retention test and (b) charge retention data for various write and erase pulses.

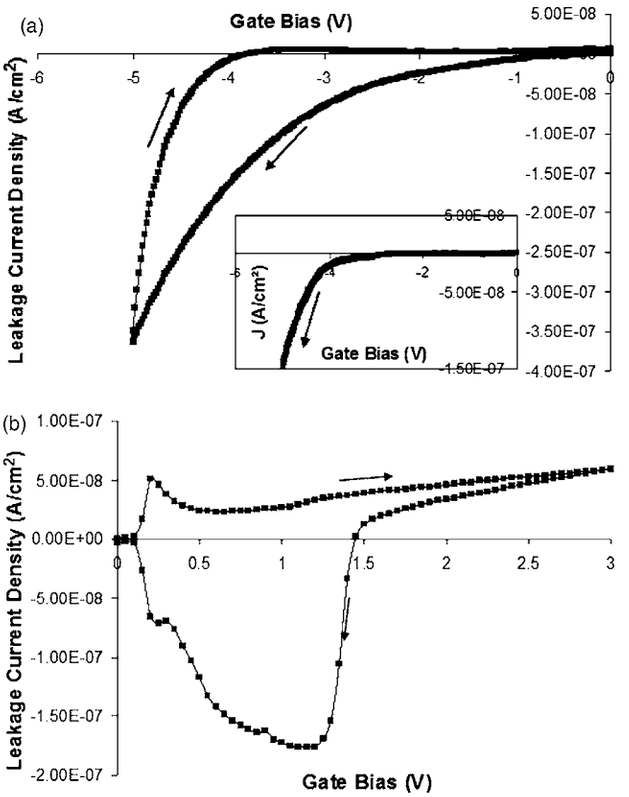


Figure 8. (a) *J-V* hysteresis of sweep voltage from 0 to −5 V and back to 0 V. Inset: 0 to −5 V sweep on previously written capacitor and (b) *J-V* hysteresis for sweep from 0 to +3 V and back to 0 V. Peaks represent charging in forward sweep and discharging in reverse sweep.

improvement in the memory window, but it only adds a small increase of the V_{FB}. The curve 4 pulse was an erase at +7 V bias for 10 s. It can be seen that not only does the flatband voltage fail to reach the fresh value immediately after the pulse, but it also drifts further away with time. This indicates that some of the trapped positive charge is very deeply trapped so that erasing efficiency at low bias is poor. Even without full erasure, extrapolations of the 10 s write and 10 s erase results to the fitted logarithmic curves lead to a large 96 mV memory window at 10^8 s.

Charge carrier transfer characteristics.— The capacitor's leakage current density vs voltage (*J-V*) hysteresis curves can be used to investigate the charge transfer characteristics. Figure 8a shows that in the negative gate bias voltage range, a large *J-V* hysteresis was detected. When the gate voltage is swept from 0 to −5 V, the leakage current is strongly dependent on the voltage. However, when the voltage is swept from −5 to 0 V, the leakage current decreases drastically first, but becomes near zero and even a small positive number that shows little bias dependence. In the first sweep stage, i.e., 0 to −5 V, the amount of holes injected from the p-type wafer increases with the magnitude of the voltage. At −5 V, a large number of holes were trapped to the embedded ITO site. In the second stage, i.e., −5 to 0 V, the injection of positive charges dropped drastically with the decrease of the magnitude of the gate bias voltages because the larger number of trapped charges screened further injection of holes from the wafer. The *J-V* hysteresis phenomenon is consistent with the *C-V* curves that show a negative shift in the V_{FB} after writing with a negative voltage. In order to verify the above positive charge trapping phenomenon, after the second sweep stage the capacitor was biased with a gate voltage from 0 to −5 V. The result is shown in the inset of Fig. 8a. The leakage current is much lower than that of the first sweep stage until the bias passes −4 V, beyond which the current begins to show a very steep slope. This is most likely due to the initiation of electrons tunneling from the low work function aluminum gate to both the ITO nanolayer and the silicon substrate. Since the *C-V* hysteresis has shown to be bias dependent, this electron flow would provide a limitation to the amount of positive charge that could be stored in the ITO site. In order to reduce the electron injection from the gate electrode, a high work function metal must be used as the electrode material.

There are several possible mechanisms that could cause the addition of positive charges to the ITO embedded capacitor during the negative write cycle. One possibility is that holes are introduced to the ITO valence band from the p-type substrate accumulation layer. However, since ITO has a larger bandgap and larger work function than those of silicon,[18] the valence band offset at the ITO/high-*k* interface should be much lower than that at the silicon/high-*k* interface. The hole retention at the low sensing biases would be energetically unfavorable. A more likely scenario would be a band-to-band type mechanism where the substrate holes tunnel directly to the conduction band of the ITO. Such a mechanism is taken from the reverse concept of applications such as organic light-emitting diodes, where the ITO conduction band is used as a hole injector into the valence band of the light-emitting material.[31] On the contrary, it is possible that a similar process occurs where the apparent hole tunneling is the result of electrons tunneling from the ITO conduction band to the silicon valence band, thus resulting in a positive charge in the ITO site. Unlike the undoped nanocrystalline silicon or germanium where minimal intrinsic charge concentration is available for tunneling, the ITO has a large concentration of free electrons[18] available for conduction, which could cause the observed effect. The third possibility is due to trap sites in the ITO or at the ITO/high-*k* interface. If the trap aligns within the substrate bandgap

during normal operation conditions the trapped charge will need to be thermally activated to the conduction or valence band for transport, or could directly tunnel to interface states at the substrate/insulator interface. Such traps may be responsible for long retention times in undoped semiconductor nanocrystals.[32]

Although the *C-V* and retention results indicate that very little charge trapping occurs on fresh samples at positive bias, the positive bias *J-V* curves display interesting features, as shown in Fig. 8b, that need to be explained for a proper understanding of the device. On the forward sweep, i.e., from 0 to 3 V, a current peak appears near 200 mV gate bias and is followed by a negative-differential-resistance dip before increasing again. A similar peak is often seen in nanocrystal memory devices.[33-35] As this effect is most commonly attributed to the Coulomb blockade effect, the fact that such a peak occurs in our structure may provide further indication that the ITO nanolayer has, in fact, been separated into discrete nanocrystals. For nanometer-scale structures, the charging energy of adding a single electron can cause an appreciable electrostatic potential change due to the nanocrystal's self-capacitance,[36] so that a sharp drop in current should occur with the addition of each electron.

On the return *I-V* sweep, i.e., from 3 to 0 V, there exists a significant dip into negative current regions as the current decreases followed by a return to low level values near zero bias. This negative current dip provides an answer to why the embedded ITO nanolayer does not display charge trapping characteristics in the *C-V* and retention measurements at positive bias. Although many nanocrystal memories operate by "writing" electrons through direct tunneling at positive bias, it has been shown that trap sites in the nanocrystal/oxide interface may be the key for the charge retention of such devices.[32] Electrons trapped in the conduction band of a nanocrystal will have enhanced potential energy due to quantum confinement and Coulomb blockade effects, creating an energetically favorable condition to tunnel back to the substrate conduction band. However, an electron trapped in a subconduction band energy level of the nanocrystal that lines up with the bandgap of the semiconductor would need thermal activation to achieve high enough energy to tunnel directly or would require a trap-assisted tunneling process to the substrate. One possible explanation for the lack of electron retention in the ITO nanolayer would be a low subconduction band trap density, so that as the positive bias returns to zero the high-energy ITO conduction band electrons that were trapped at high bias are "dumped" back to the silicon conduction band by direct tunneling. Since the ITO is degenerately n-type, even if these traps do exist they would be below the Fermi level so that the states would be filled and inaccessible to added electrons. Also, localized leakage paths could cause quick discharge of the entire floating gate area if the ITO forms a continuous or partially continuous nanolayer. Although the ITO is in the form of nanocrystals, isolated leakage currents can still discharge the entire floating gate if the aerial coverage is greater than 50% because lateral tunneling between dots begins to dominate over localized charge storage.[36] Lastly, it is likely that the poor electron retention in comparison to hole retention could be due to the difference in barrier heights for electrons and holes in the Zr-doped HfO_2. The barrier height for holes is over twice as large as that for electrons (approximately 3.2 eV compared to 1.5 eV for Si conduction band and valence band, respectively).[26] As such, the transparency of the barrier as seen by electrons will be significantly greater than that seen by holes. In this case, the band offset mismatch may create an opportunity to achieve ultrathin tunnel oxides with excellent retention if holes, instead of electrons, are targeted as the memory carriers.

Conclusion

In this paper, it has been shown that an ITO-embedded zirconium-doped hafnium oxide MOS can exhibit promising nonvolatile memory characteristics through positive charge trapping during negative write cycles. The ability to trap and retain these charges has been shown to have high dependency on control and tunnel high-*k* dielectric thicknesses. Extrapolations of retention data indicate a memory window of near 100 mV after 10^8 s. Negative differential resistance in the *J-V* curve indicates electron tunneling and Coulomb blockade effects at positive bias, but retention of these negative charges is poor in comparison to the holes trapped at negative bias. The asymmetrical band offsets of Zr-doped HfO_2 may allow for hole retention at thinner tunnel oxide thicknesses where electron retention is negligible. Further improvement of charge trapping and retention may be achieved by replacing the aluminum gate with a large work function gate electrode material to reduce electron injection from the gate, and varying oxide thicknesses to improve writing and erasing efficiency and data retention.

Acknowledgments

The authors acknowledge the NSF DMI-0300032 grant for partial support of this research. Dr. J. Lu is acknowledged for conducting X-ray diffraction experiments for the samples in this study, and his guidance and valuable technical discussions during the preparation of this manuscript. One of the authors, A.B., acknowledges financial support from Applied Materials, Inc.

Texas A&M University assisted in meeting the publication costs of this article.

References

1. P. Cappellitti, C. Golla, P. Olivo, and E. Zanoni, *Flash Memories*, p. 4, Kluwer, Boston (1999).
2. P. Pavan, R. Bez, P. Olivo, and E. Zanoni, *Proc. IEEE*, **85**, 1248 (1997).
3. R. Bez, E. Camerlenghi, A. Modelli, and A. Visconti, *Proc. IEEE*, **91**, 489 (2003).
4. E. H. Nicollian and J. R. Brews, *Metal Oxide Semiconductor Science and Technology*, p. 478, Wiley, Hoboken, NJ (2003).
5. D. Schroder, *Semiconductor Material and Device Characterization*, p. 261, Wiley-Interscience, New York (1990).
6. S. Mori, Y. Yamaguchi, M. Sato, H. Meguro, H. Tsunoda, E. Kamiya, K. Yoshikawa, N. Arai, and E. Sakagami, *IEEE Trans. Electron Devices*, **43**, 47 (1996).
7. N. Mielke and J. Chen, in *Oxide Reliability: A Summary of Silicon Oxide Wearout, Breakdown, and Reliability*, D. J. Dumin, Editor, p. 103, World Scientific, Singapore (2002).
8. B. De Salvo, G. Ghibaudo, G. Pananakakis, P. Masson, T. Baron, N. Buffet, A. Fernandes, and B. Guillaumot, *IEEE Trans. Electron Devices*, **48**, 1789 (2001).
9. C. Li, W. Fan, B. Lei, D. Zhang, S. Han, T. Tang, X. Liu, Z. Liu, S. Asano, M. Meyyappan, J. Han, and C. Zhou, *Appl. Phys. Lett.*, **84**, 1949 (2004).
10. S. Tiwari, F. Rona, K. Chan, L. Shi, and H. Hanafi, *Appl. Phys. Lett.*, **68**, 1377 (1996).
11. J. De Blauwe, *IEEE Trans. Nanotechnol.*, **1**, 72 (2002).
12. W. Choi, W. Chim, C. Heng, L. Teo, V. Ho, V. Ng, D. Antoniadis, and E. Fitzgerald, *Appl. Phys. Lett.*, **80**, 2014 (2002).
13. D. Zhao, Y. Zhu, R. Li, and J. Liu, *Solid-State Electron.*, **50**, 362 (2006).
14. Z. Liu, C. Lee, V. Narayanan, G. Pei, and E. C. Kan, *IEEE Trans. Electron Devices*, **49**, 1606 (2002).
15. J. Lee, Y. Harada, J. Pyun, and D.-L. Kwong, *Appl. Phys. Lett.*, **86**, 103505 (2005).
16. C. Lee, U. Ganguly, V. Narayanan, T.-H. Hou, J. Kim, and E. Kan, *IEEE Electron Device Lett.*, **26**, 879 (2005).
17. Y. Park, V. Choong, Y. Gao, B. Hsieh, and C. Tang, *Appl. Phys. Lett.*, **68**, 19 (1996).
18. O. Malik, V. Grimalsky, and J. De la Hidalga, *J. Non-Cryst. Solids*, **352**, 1461 (2006).
19. G. Wilk, R. Wallace, and J. Anthony, *J. Appl. Phys.*, **87**, 484 (2000).
20. J. Robertson, *Rep. Prog. Phys.*, **69**, 327 (2006).
21. J. Lee, X. Wang, W. Bai, N. Lu, and D.-L. Kwong, *IEEE Trans. Electron Devices*, **50**, 2067 (2003).
22. J. Lu and Y. Kuo, *Appl. Phys. Lett.*, **87**, 232906 (2005).
23. Y. Kuo, *ECS. Trans.*, **3**(3), 253 (2006).
24. J. Lu, Y. Kuo, and J.-Y. Tewg, *J. Electrochem. Soc.*, **153**, G410 (2006).
25. J.-Y. Tewg, Y. Kuo, and J. Lu, *Electrochem. Solid-State Lett.*, **8**, G27 (2005).
26. Y. Kuo, J. Lu, S. Chatterjee, J. Yan, T. Yuan, H.-C. Kim, W. Luo, J. Peterson, and M. Gardner, *ECS. Trans.*, **1**, 447 (2006).
27. D. Triyoso, *ECS. Trans.*, **3**(3), 463 (2006).
28. J. Hauser and K. Ahmed, *Characterization and Metrology for ULSI Technology*, p. 235, AIP, New York, (1998).
29. D. D. L. Chung, P. W. De Haven, H. Arnold, and D. Ghosh, *X-Ray Diffraction at Elevated Temperatures: A Method for In-Situ Process Analysis*, Chap. 1, VCH, New York (1993).
30. S. Hu, D. Kerr, and L. Gregor, *Appl. Phys. Lett.*, **10**, 97 (1967).
31. R. Friend, R. Gymer, A. Homes, J. Burroughes, R. Marks, C. Taliani, D. Bradley, D. Dos Santos, J. Bredas, M. Logdlund, and W. Salaneck, *Nature (London)*, **397**, 121 (1999).
32. Y. Shi, K. Saito, H. Ishikuro, and T. Hiramoto, *J. Appl. Phys.*, **84**, 2358 (1998).

33. T. Maeda, E. Suzuki, I. Sakata, M. Yamanaka, and K. Ishii, *Nanotechnology*, **10**, 127 (1999).
34. M. Shalchian, J. Grisolia, G. Ben Assayag, H. Coffin, S. M. Atarodi, and A. Claverie, *Solid-State Electron.*, **49**, 1198 (2005).
35. J. Lu, Y. Kuo, J. Yan, and C.-H. Lin, *Jpn. J. Appl. Phys., Part 2*, **45**, L901 (2006).
36. R. Rao, R. Steimle, M. Sadd, C. Swift, B. Hradsky, S. Straub, T. Merchant, M. Stoker, S. Anderson, M. Rossow, J. Yater, B. Acred, K. Harber, E. Prinz, B. White, Jr., and R. Muralidhar, *Solid-State Electron.*, **48**, 1463 (2004).

ECS Transactions, 3 (3) 253-263 (2006)
10.1149/1.2355717, copyright The Electrochemical Society

Mixed Oxide High-K Gate Dielectrics - Interface Layer Structure, Breakdown Mechanism, and New Memories

Yue Kuo

The Thin Film Nano and Microelectronics Research Laboratory, Texas A&M University, College station, TX, 77843-3122, USA

Mixed metal oxide high-k films have many advantageous dielectric properties over conventional metal oxides. In this paper, three major subjects on the mixed oxide gate stack prepared by the sputtering method, i.e., dopant influence on interface layer structures, double-layer gate stack breakdown mechanism, and new CMOS-type nonvolatile memories based on nanocrystals embedded doped oxide high-k films, are examined and discussed.

Advantages of Mixed Oxide High-K Gate Dielectrics

According to the Semiconductor Technology Roadmap, the thermally grown SiO_2 gate dielectric needs to be replaced by a high-k thin film for future generation MOSFETs. The sub 1.2nm thick SiO_2 film has many practical problems, such as a high leakage current, undesirable dopant diffusion from the gate, and difficulty in film thickness control. In principle, a high-k dielectric can avoid these problems because a thicker layer can be used.

Metal oxides, such as HfO_2, ZrO_2, and Ta_2O_5, are promising high-k gate dielectric candidates. They are easy to prepare by methods such as CVD, MOCVD, ALD, and sputtering. The film is easily annealed in a tube furnace or a rapid thermal annealing equipment. However, devices made of this kind of high-k material are inferior than expected, e.g., low effective k values, high leakage currents, and large densities of interface states. These problems are originated from some undesirable material properties, such as crystalline phase formation, uncontrollable interface reactions, small band gaps, and low electron offsets with silicon. Theses problems are caused by the thermodynamic nature of the material. They can only be partially solved by adjusting process conditions, such as lowering the deposition temperature, shortening the annealing time, avoiding short wavelength light exposure or ion bombardment, or annealing under low oxygen concentration atmosphere. The fundamental way to solve these problems is to get around the thermodynamic barrier, for example, by changing the material composition or structure. Many research groups have demonstrated that physical, chemical, and electrical properties of the high-k material can be greatly improved by doping it with a third element (1-11). The doped oxide is also called the mixed oxide because the final film contains dopant atoms in the oxidized form. Excellent results have been obtained from this approach. For example, the crystallization of Ta_2O_5 was suppressed by doping with Si, Al, or Zr (6,7). The high-k/Si interface layer thickness was thinned after doping (5). The band gap energy was increased with the increase of the dopant concentration (4). The interface density of states was minimized with the

inclusion of a proper amount of dopant (8). In recent years, research activities on mixed oxide high-k materials increased drastically because of these advantages (3).

In this paper, the author examines three critical areas in the mixed oxide high-k dielectric, i.e., interface structure, breakdown process, and memory applications. New results and detailed discussions are presented.

Dopant Influence on Interface Layer Structures

The dopant atom influences the bulk high-k film's physical and chemical properties, such as binding energies, bond structures, etc. These characteristics directly affect the film's electrical characteristics, such as the k value, leakage current, breakdown strength, relaxation current, etc. For example, the binding energies of Ta and O in Ta_2O_5 shift to lower values with the addition of Hf or Zr dopants (5,10). This is because that both Hf and Zr have lower electronegativities, i.e., 1.30 and 1.33 eV separately, than those of Ta and O, i.e., 1.50 and 3.44 eV separately. The band gap of Ta_2O_5 increases with the Hf dopant content from the O1s energy loss spectra and the internal photoemission spectra (4,12).

Since the interface layer formation is usually unavoidable, as the bulk high-k film thickness decreases, the influence of the interface layer on the gate stack electrical properties becomes more pronounced. For example, for a gate stack with an equivalent oxide thickness (EOT) near 1 nm, the physical thickness of the interface layer can be the same as or even larger than that of the bulk film. Since the interface layer often has inferior dielectric properties than the bulk layer, it can dominate the final gate stack properties. When the high-k film is doped, its interface layer composition and structure are changed through two routes: interface reactions and the mechanism of reactant transport through the bulk film. For example, in many cases, the dopant atom can be more reactive than the metal atom in the original high-k film when in contact with the adjacent silicon substrate. Figure 1 shows Si 2p binding energies of (a) the undoped Ta_2O_5/silicon interface layer and (b) the Hf-doped Ta_2O_5/silicon interface layer (4). Ta does not exist in the former interface layer, but Hf exists in the latter interface layer. This result is consistent with the SIMS profile result (5). The same phenomena exist in the Zr-doped Ta_2O_5/silicon interface layer (10).

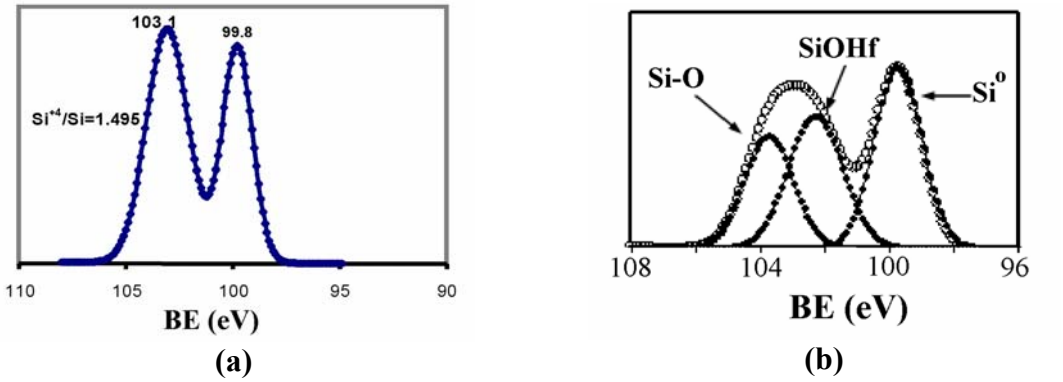
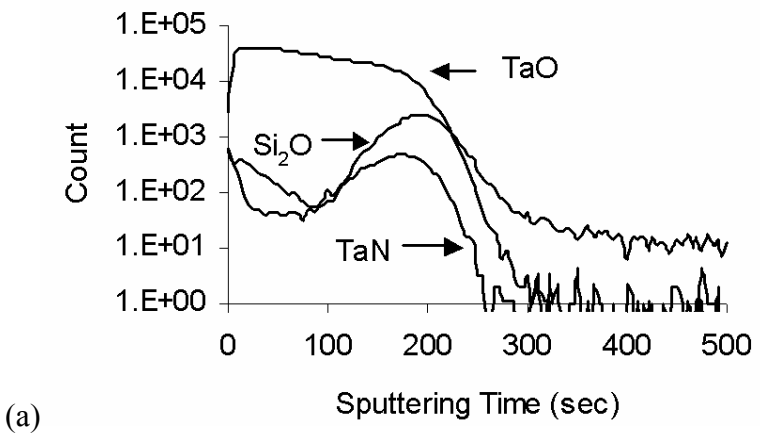
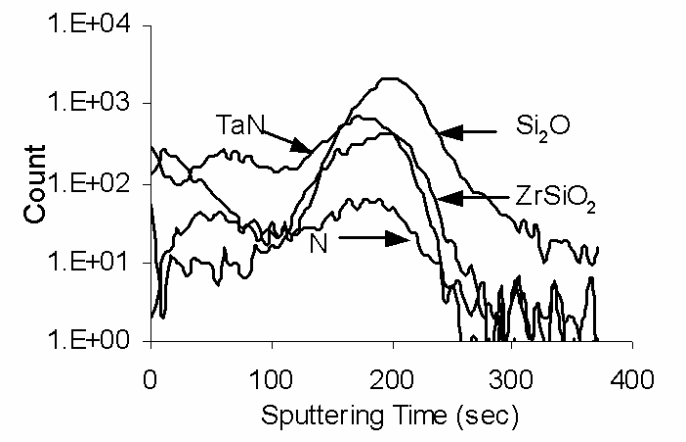


Figure 1. Si 2p at (a) undoped Ta_2O_5 and (b) Hf-doped Ta_2O_5/silicon interfaces (4).

The dopant atoms can penetrate an interface layer inserted between the high-k film and the silicon substrate. Figure 2 shows SIMS profiles of (a) $TaO_x/TaN_x/Si$ and (b) Zr-doped $TaO_x/TaN_x/Si$ structures (13). The Zr dopant migrates through the TaNx interface layer and forms a $ZrSiO_2$ component in the new interface layer. The same phenomenon occurs when Hf is the dopant in the TaO_x high-k film. When the inserted TaN_x interface layer is replaced by a thin SiO_2 or SiON layer, similar type of silicates were detected at the interface. In addition, the interface of Fig. 12(b) sample also includes the TaO_xN_y component, which shows the complication of the dopant involvement (13).



(a)



(b)

Figure 2. SIMS profiles of (a) $TaO_x/TaN_x/Si$ and (b) Zr-doped $TaO_x/TaN_x/Si$ structures (13).

The silicate-type interface layer has a larger k value than that of SiO_2 or SiON. Therefore, the EOT of the doped high-k gate stack can be lower than that of the undoped high-k stack because of the formation of the new interface layer (8). The interface density of states (D_{it}) of the doped high-k film is lower than that of the undoped high-k film, as shown in Figure 3(a) (8). The Zr dopant concentration is controlled by the target co-sputtering power, i.e., 0, 15, or 24W, while the Hf sputtering power was fixed at 60W. All films have an EOT < 2 nm except undoped HfO_2 at 2.8 nm. The high-k film with N in the interface layer has a higher D_{it} than that without N.

Furthermore, the diffusion of O_2 through the bulk high-k film can be hindered with the existence of certain type of dopants, such as Hf in Ta_2O_5 (5). Figure 4 shows that the interface layer thickness measured from TEM pictures decreases with the content of

Hf dopant in the high-k film (5). At the early stage, the interface layer is instantaneously formed due to the quick interface reaction. With the increase of the annealing time, the interface layer growth is controlled by the O diffusion rate through the bulk high-k film. Eventually, the interface layer thickness reaches a steady state value because of the low diffusion rate. The high Hf concentration in the bulk film slows the diffusion process, which makes the interface layer thinner than that of the low Hf concentration film.

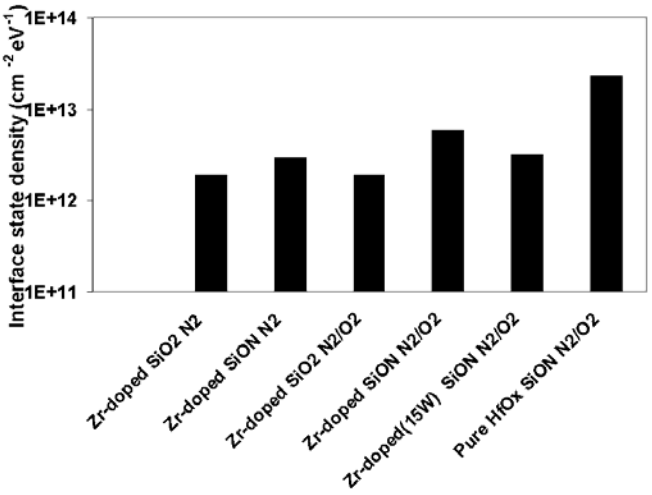


Figure 3. D_{it}'s of undoped and Zr-doped HfO_x with SiO_2 or SiON interface (8).

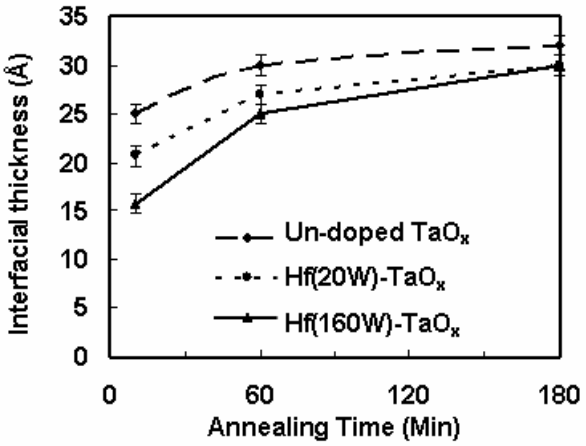


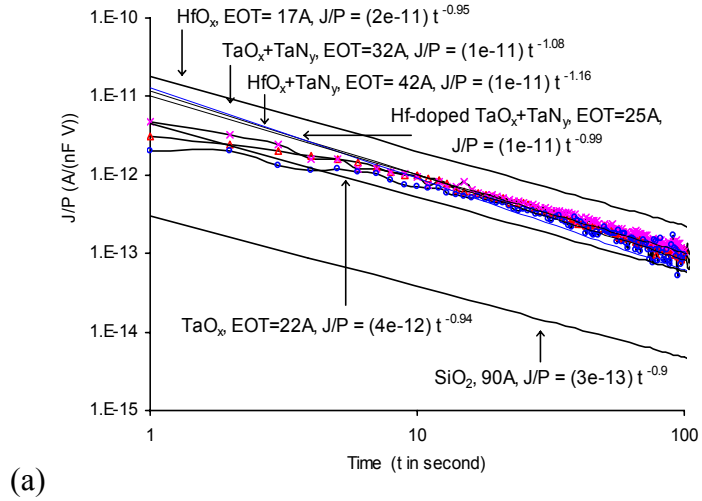
Figure 4. Interface layer thickness vs. annealing time at 700°C O_2 (5).

Most dielectric properties of the high-k gate stack film, such as EOT, D_{it}, fixed charge density (Q_f), leakage current, and breakdown strength, are greatly improved due to the change of interface structure. .

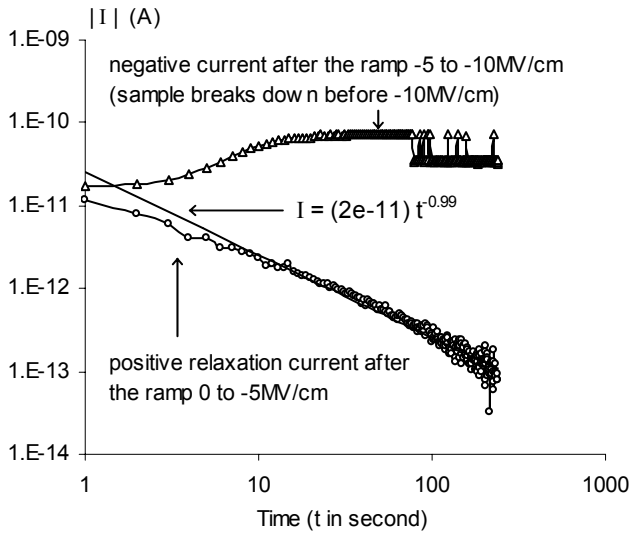
Double-Layer Gate Stack Breakdown Mechanism

Since the interface layer formation between a high-k film and the silicon wafer is unavoidable, the breakdown process is influenced by both the bulk and the interface layers. As discussed before, the dopant can alter the film's bond structure, which is directly related to its polarization characteristics and breakdown mechanism.

Dielectric relaxation current is a bulk-related phenomenon, which occurs upon the sudden removal or application of a voltage (V). It follows the direction of the dV/dt and is time-dependent. Relaxation current is a function of electric polarization, and it decays following the Curie-von Schweidler law $J/P=at^{-n}$, where J is the relaxation current density (A/cm^2), P is the total polarization or surface charge density ($V \cdot nF/cm^2$), t is time in seconds, a is a constant, and n is a real number close to 1 (14). Due to its high polarizable bond structure, the metal oxide high-k film usually has a large relaxation current compared with the SiO_2 film. Figure 5(a) shows relaxation current vs. time of undoped and Hf-doped Ta_2O_5 high-k stacks with an inserted TaN_x interface layer. The relaxation currents are measured after the sudden removal of a constant voltage on the gate. The undoped and doped films have similar relaxations because Hf-O and Ta-O have similar polarizability. Their relaxation currents are one to two orders of magnitude larger than that of SiO_2. The relaxation current disappears if the film breaks, which can be detected from the change of direction of current flow, as shown in Figure 5(b) (15).



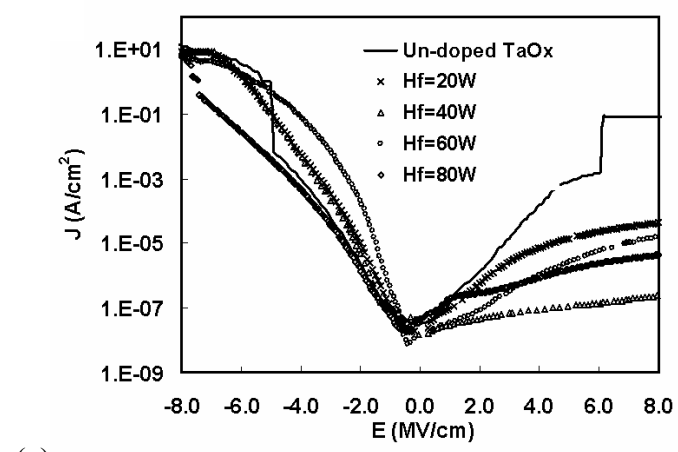
(a)



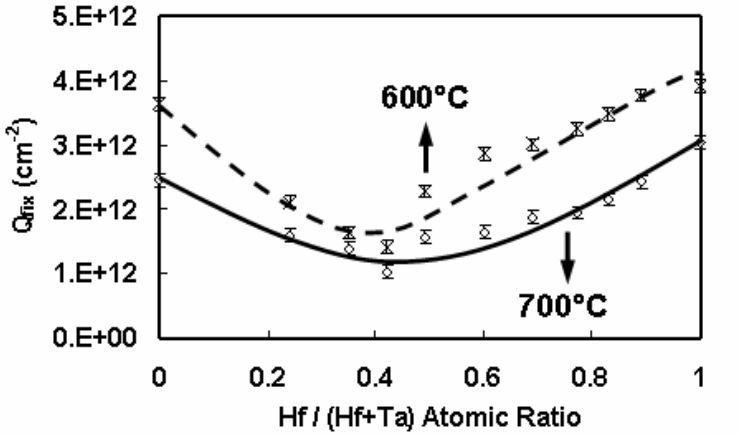
(b)

Figure 5. (a) Relaxation current vs. leakage current. (b) relaxation currents of undoped and Hf-doped Ta_2O_5 with an inserted TaN_x interface layer (15).

The dopant concentration affects the gate stack breakdown strength. Figure 6(a) shows J-E curves of undoped and lightly Hf-doped Ta_2O_5 high-k stacks (5). The undoped film has a breakdown strength of -4.5 mV/cm in the accumulation region and a breakdown strength of 6 mV/cm in the inversion region while films with Hf sputtering power < 80W do not break in the test voltage range. However, the film with the 80W Hf sputtering power has a breakdown strength close to 8 mV/cm. Separately, it was observed that the film's breakdown strength decreased when the Hf sputtering power was higher than 80W. This relationship is consistent with the high-k stack's fixed charge density (Q_f), i.e., the low Q_f sample has a high breakdown strength, as shown in Figure 6(b) (5).



(a)



(b)

Figure 6. (a) J-E curves of undoped and Hf-doped TaO_x. Hf sputtering power 20~80W ~ Hf/(Hf+Ta) 0.25~0.49, (b) Q_f vs. Hf/(Hf+Ta) after 600ºC and 700ºC O_2 annealing (5).

The high-k stack typically shows one- or two-step breakdown mode. They are originated from the same mechanism and determined by the relative strengths of individual layers. The dopant type and concentration are key factors of the strength. In spite of the different measurement methods, e.g., J-V, constant voltage stress, or time dependent stress, the breakdown results are consistent. The breakdown sequence of the high-k stack, i.e., whichever layer breaks first, can be judged from the relaxation current.

The breakdown of a high-k stack can be evaluated with the ramp-relax test using a MOS capacitor. The process includes two steps: first, a negative ramping gate voltage V_{ramp} is applied and the leakage current J_{ramp} is measured; second, the gate bias voltage is

switched to a very low monitor voltage V_m, such as −0.01, 0, or +0.01 V, for a short period, such as half a second, to measure the monitor current J_m. If the gate dielectric layer is composed of SiO_2 only, before breakdown, the polarity of J_m varies with the polarity of V_m. However, when the gate dielectric layer is composed of a high-k stack, such as Zr-doped HfO_x/SiO_2, before breakdown, the polarity of J_m is independent of the polarity of V_m because the Zr-doped HfO_x layer has a large relaxation current (16).

The breakdown sequence of a double-layer dielectric stack can be delineated with the above method. For example, Figure 7 shows the J_{ramp} and J_m vs. V_{ramp} curves of two gate stacks, i.e. (a) TiN gate/Zr-doped $HfO_x/SiO_2/p$-Si and (b) Al gate/Hf-doped TaO_x/ silicate/ p-Si (16). For the former, J_m changes at the second jump of J_{ramp}; for the latter, J_m changes at the first jump of J_{ramp}. The Fig. 7(b) sample wafer has a lower dopant concentration than the Fig. 7(a) sample wafer, which is the cause of its lower J_m after breakdown. The failure of Fig. 7(a) sample is initiated from the SiO_2 interface breakdown because the J_{ramp} and J_m are of opposite polarities after first breakdown. For Fig. 7(b) sample, the interface silicate layer is a type of high-k film. The breakdown started from the bulk high-k layer because the high stress voltage created additional traps in Hf-doped TaO_x film and induced hole trapping near the gate side. Therefore, the breakdown sequence of the high-k stack can be estimated from its relaxation behavior.

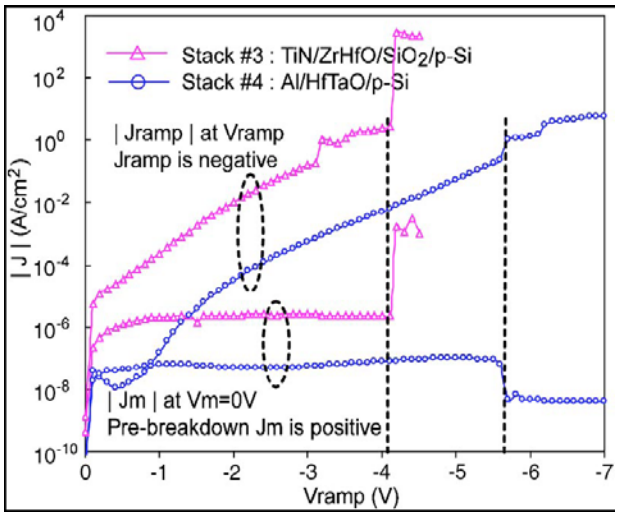


Figure 7. Two-step breakdown mode of TiN/Zr-doped HfO_x/1 nm SiO_2/p-Si and Al/Hf-doped TaO_x/silicate/p-Si from ramp-relax tests (16).

New CMOS-Type Nonvolatile Memories Based on Nanocrystals Embedded Doped Oxide High-K Films

The floating gate device is a nonvolatile memory. It is usually composed of a gate dielectric film embedded with a polysilicon layer. This kind of device is difficult to shrink into very small geometry without serious charge loss issues. When the embedded polysilicon layer is replaced by a nanocrystalline silicon (nc-Si) layer, electrons can be stored in discrete states on nc-Si sites. Due to high barrier height between nc-Si and SiO_2, the charge transfer among nc-Si sites is greatly reduced and the retention time can be extended. However, in order to save the operation power, the tunneling SiO_2, which is located between the nc-Si layer and the silicon substrate, needs to be very thin, such as <

2 nm. The charge retention capability becomes an issue again. This problem can be solved by replacing the SiO_2 layer with a high-k material, which can has a lower EOT but a larger physical thickness than SiO_2. However, the conventional metal oxide high-k film has a low crystallization temperature and a low electron offset with silicon. If the doped metal oxide high-k dielectric is used to substitute the conventional metal oxide high-k film, the above problems can be avoided.

We have successfully fabricated two types of MOS capacitors with nanocrystals embedded in Zr-doped HfO_2 high-k dielectric layer (17,18,19). After the stacked high-k layers were sequentially deposited by sputtering without breaking the vacuum, a rapid thermal annealing step was carried out at 950°C for 1 minute under N_2:O_2 (1:1). The embedded ITO was crystallized from the x-ray diffraction (XRD) measurement (17). Figure 8(a) shows C-V curves of an ITO embedded capacitor measured by sweeping the gate voltage from negative to positive values, i.e., (-3,3V), (-7,7V), and (-9,9V). Holes were trapped in this capacitor because all curves are on the negative side of the C-V curve of the capacitor without the embedded ITO layer, as shown in Fig. 8(c). In addition, the amount of holes trapped in the capacitor increased with the magnitude of the gate voltage, i.e., V_{FB} of –0.46, -0.57, -0.74, and –1.1V vs. (-3,3V), (-5,5V), (-7,7V), and (-9,9V), respectively. Figures 8(b) shows C-V hysteresis curves of Fig. 8(a) sample swept from –9V to 9V ("forward") and then back to –9V ("backward") (19). It has a counterclockwise flat band voltage shift (ΔV_{FB}) of 0.56V. Its current-voltage (J-V) curve contains a NDR peak similar to that of a floating gate memory. Figures 8(c) shows C-V hysteresis curves of the Zr-doped HfO_2 high-k film without an embedded layer. It has a very small ΔV_{FB} of 0.07V. Figures 8(d) shows C-V hysteresis curves of a nc-Si embedded sample, which has a ΔV_{FB} of 0.32V.

There are major differences between the ITO embedded and nc-Si embedded samples. First, the ITO embedded sample has negative V_{FB} values from both "forward" and "backward" sweep curves, i.e., holes trapped in the accumulation region were only partially neutralized by electrons tunneled from the substrate in the inversion region. Therefore, after the "forward" and "backward" sweep cycle, the net charge is positive. However, the nc-Si embedded sample has a negative V_{FB} (-0.49V) in the "forward" direction and a positive V_{FB} (0.17V) in the "backward" direction. There are more electrons tunneled from the substrate in the inversion region than necessary to neutralize holes tunneled from the substrate in the accumulation region. The net charge in this capacitor is negative. Second, the ITO embedded sample has a negative gate voltage corresponding to the J-V curve's NDR peak while the nc-Si embedded sample has a positive gate voltage corresponding to the NDR peak. This further confirms the hole trapping characteristics of the ITO embedded sample. Separately, it has been observed that the ΔV_{FB} of the hysteresis curves of Fig. 8(b) sample increased with the increase of the sweep voltage range. Therefore, the amount of holes trapped in the ITO embedded sample can be adjusted by varying the applied gate voltage.

Holes are deeply trapped in the ITO embedded high-k film with a large "read" to "erase" window, e.g., 0.15V. Since the trapped holes are difficult to tunnel back to the substrate, the device has good charge retention. However, more studies are required to completely detrap the strongly held holes. The availability of electrons and holes "trapping" and "detrapping" capability makes it possible to fabricate CMOS-type memories, which broadens the function and design of IC products.

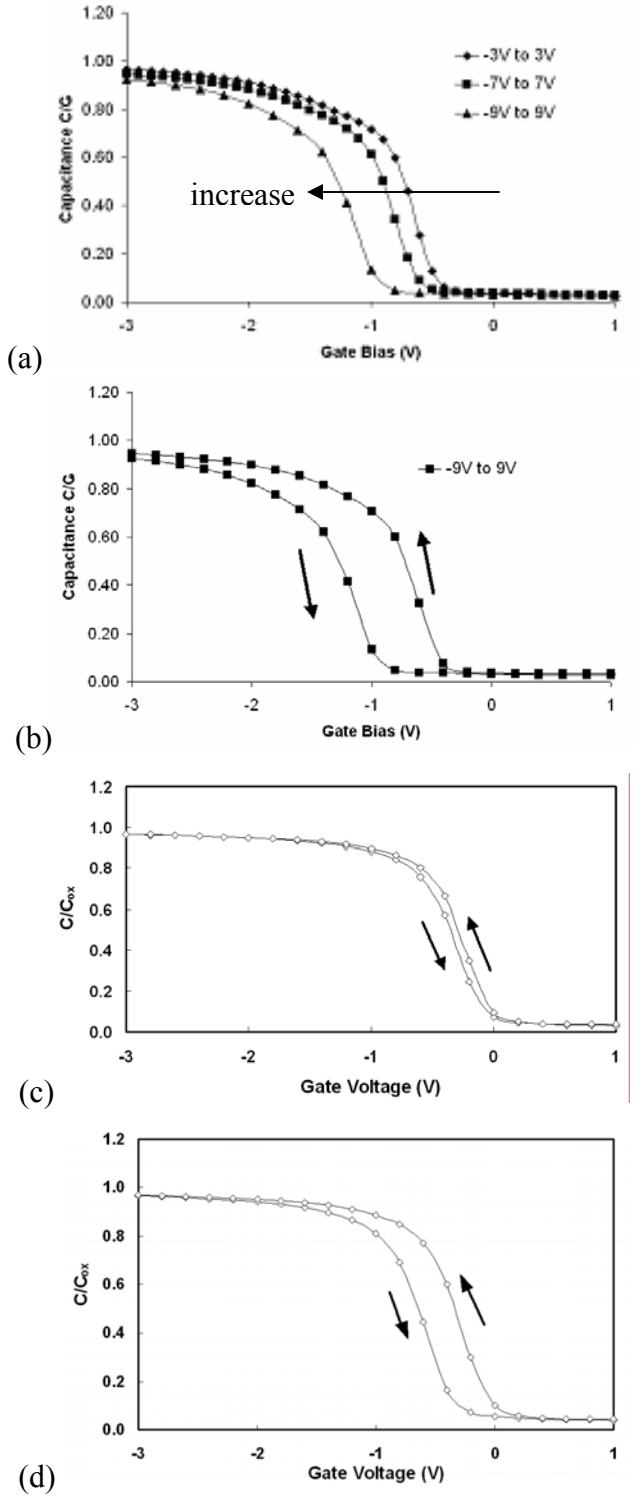


Figure 8. (a) C-V curves of an ITO embedded capacitor at different gate sweep voltages, (b) C-V hysteresis curves of (a) sample, (c) C-V hysteresis curve of sample without an embedded layer, and (d) C-V hysteresis curves of nc-Si embedded capacitor. Samples (b), (c), and (d) were swept between –9 and 9V.

Summary

The doped oxide high-k films have many superior dielectric properties than the undoped oxide film. In this paper, the author reviewed three major subjects on the doped oxide. First, the interface layer composition and structure are drastically changed by the existence of the dopant. Second, the high-k gate stack breakdown process is improved by the existence of the dopant. The breakdown sequence can be determined by the pattern of the relaxation current change. Third, the doped high k film can be used in the nonvolatile memory. Electron and hole memory devices were fabricated by embedding nc-Si and ITO layers, separately, which opens many opportunities for new applications.

Acknowledgments

This work is partially supported by the NSF project DMII-0429176. Author would like to acknowledge his graduate students and postdoctoral scholars J.-Y. Tewg, J. Lu, W. Luo, J. Yan, T. Yuan, A. Birge, C.-H. Lin, S. Chatterjee for their excellent work. He also thanks Prof. W. Kuo of University of Tennessee for technical discussions and J. Peterson of SEMATECH for many useful suggestions and wafer supplies.

References

1. L. Manchanda, M. D. Morris, M. L. Green, R. B. van Dover, F. Klemens, T. W. Sorsch, P. J. Silverman, G. Wilk, B. Busch, and S. Aravamudhan, *Microelectron. Eng.*, **59**, 351 (2001).
2. C. Zhao, T. Witters, B. Brijs, H. Bender, O. Richard, M. Caymax, T. Heeg, J. Schubert, V. V. Afanas'ev, A. Stesmans, and D. G. Schlom, *Appl. Phys. Lett.*, **86**, 132903 (2005).
3. Y. Kuo, *Electrochem. Soc. Trans.*, **2**(1), 13 (2006).
4. J. Lu and Y. Kuo, *Appl. Phys. Lett.*, **87**(23), 232906 (2005).
5. J. Lu, Y. Kuo, and J.-Y. Tewg, *J. Electrochem. Soc.,* **153**, G410 (2006).
6. Y. Kuo, J.-Y. Tewg, and J. P. Donnelly, *ECS Proc. Intl. Semiconductor Technology Conf*., **2001-17**, 324 (2001).
7. J.-Y. Tewg,Y. Kuo, and J. Lu, *Electrochem. Solid-State Letts.*, **8**(1), G27 (2005).
8. Y. Kuo, J. Lu, S. Chatterjee, J. Yan, H. C. Kim, T. Yuan, W. Luo, J. Peterson, and M. Gardner, *ECS Trans. High Dielectric Constant Gate Stacks III*, 548 (2005).
9. J.-Y. Tewg, Y. Kuo, J. Lu, and B. W. Schueler, *J*. Electrochem. Soc., **152**(8), G617 (2005).
10. J.-Y. Tewg, Y. Kuo, J. Lu, and B. W. Schueler, *J. Electrochem. Soc.*, **151**(3), F59 (2004).
11. J. Lu, Y. Kuo, J. -Y. Tewg, and B. Schueler, *Vacuum*, **74**(3-4), 539 (2004).
12. V. V. Afanas'ev, A. Stesmans, C. Zhao, M. Caymax, Z. M. Rittersma, and J. W. *Maes, Appl. Phys. Lett.***, 86**, 072108 (2005).
13. J. Lu, Y. Kuo, J.-Y. Tewg, and B. Schueler, *Vacuum*, **74**(3-4), 539 (2004).
14. A. K. Jonscher, *J. Physics, D: Appl. Phys.*, **32**, 57 (1999).
15. W. Luo, Y. Kuo, and W. Kuo, *IEEE Trans. Dev. Mat. Reliability*, **4**(3), 488 (2004).
16. W. Luo, T. Yuan, Y. Kuo, J. Lu, J. Yan, and Way Kuo, *Appl. Phys. Letts.,* **88**, 202904 (2006).

17.Y. Kuo, J. Lu, J. Yan, and C.-H. Lin, *6th IEEE Conference on Nanotechnology*, in press, 2006.
18. J. Lu, Y. Kuo, J. Yan, and C.-H. Lin, *Jpn. J. Appl. Phys.*, in press, 2006.
19. A. Birge, C.-H. Lin, and Y. Kuo, *ECS Trans. 4th Intl. Symp. High Dielectric Constant Gate Stacks*, submitted, 2006.

International Journal of Reliability, Quality and Safety Engineering
Vol. 15, No. 6 (2008) 515–538

World Scientific
www.worldscientific.com

MULTIOBJECTIVE DESIGN OF EQUIVALENT ACCELERATED LIFE TESTING PLANS

HAITAO LIAO*,† and ZHAOJUN LI‡

†*Department of Nuclear Engineering*
Department of Industrial and Information Engineering
University of Tennessee, Knoxville
TN 37996, USA
†*hliao4@utk.edu*

‡*Department of Industrial Engineering*
University of Washington
Seattle, WA 98195, USA

This paper is focused on the multiobjective design of equivalent accelerated life test (ALT) plans. Equivalent ALT plans are expected to achieve the same statistical performance as a baseline ALT plan yet lead to other desired performance measures such as reduced test time and total cost. Before determining the desired multiobjective equivalent ALT plans, an efficient fast non-dominated sorting genetic algorithm (NSGA-II) is utilized to identify a set of Pareto optimal solutions. To handle a large number of Pareto optimal solutions, a self-organizing map (SOM) and data envelopment analysis (DEA) are sequentially applied to classify the Pareto solutions and reduce the size of the suggested solution set. This integrated approach allows for the tradeoff of information among the Pareto solutions and the reduction in the size of the solution set. It provides a useful tool for practitioners to make meaningful decisions in planning ALT experiments.

Keywords: Equivalency; accelerated life testing plan; multiobjective decision making; NSGA-II; SOM; DEA.

Acronym

ALT	accelerated life testing
Asvar	asympototic variance
BMU	best matching unit
CDF	cumulative distribution function
DEA	data envelopment analysis
DMU	decision making units
MLE	maximum likelihood estimate
MOSO	multiple objective seletion optimization
NSGA-II	the efficient fast non-dominated genetic algorithm

*Corresponding author.

pdf probability density function
SOM self-organizing map

Notation

S_0 stress level under normal operating conditions
S_H highest stress level that could be used in an ALT experiment
k number of stress levels in an ALT experiment
S_j stress level j used in an ALT experiment, $j = 1, 2, \ldots, k$
Z_j standardized stress level j, $Z_j = (S_j - S_0)/(S_H - S_0)$
n_j total number of test units allocated to stress level j
N total number of test units $N = \sum_{j=1}^{k} n_j$
t_c censoring time in an ALT experiment
$\underline{\Theta}$ vector of parameters of an ALT model
$\widehat{\underline{\Theta}}$ MLE of $\underline{\Theta}$
$\mathbb{I}_{\widehat{\underline{\Theta}}}$ information matrix associated with an ALT plan
$\mathbb{B}^T$ transpose of matrix (or vector) $\mathbb{B}$
$\|\mathbb{B}\|$ norm of vector $\mathbb{B}$, or the absolute value of a scalar
$\underline{x}$ vector of decision variables $[x_1, x_2, \ldots, x_m]$ in planning ALT
$\mathbb{X}$ variance-covariance matrix of $\widehat{\underline{\Theta}}$ from the baseline ALT plan with decision variables $\underline{x}^{(0)}$
$\mathbb{Y}$ variance-covariance matrix of $\widehat{\underline{\Theta}}$ from the equivalent ALT plan with decision variables $\underline{x}^{(1)}$

1. Introduction

1.1. *Background*

Due to increasing global competition, today's industry expects to obtain the reliability information of new products as quick as possible before releasing the products. However, as design and manufacturing technologies become more advanced it is difficult, if not impossible, to acquire such reliability information within a short period of time under normal operating conditions. To solve the problem, accelerated lift testing (ALT) has been widely used to estimate the reliability of a product by applying severer-than-normal stress conditions (e.g., higher temperature, humidity and pressure levels) during testing. Besides, step-stress and more complex stress loadings rather than constant-stress have also been utilized to further shorten the total testing time. In conjunction with an appropriate ALT model, a well designed ALT plan, which determines experimental settings such as stress levels, unit allocation to each stress level and testing termination time, helps improve the estimation accuracy of the ALT model.

Elsayed[1] classifies the existing ALT models into three categories: statistics-based models, physics-statistics-based models, and physics-experimental-based models.

Specifically, accelerated failure time (AFT) models, which usually belong to the statistics-based models or physics-statistics-based models, have been widely used. An AFT model assumes that the stress affects the failure time multiplicatively and the life distributions under normal operating conditions and accelerated conditions belong to the same distribution family but with different location parameters. The well known Arrinehius-Weibull model, inverse power law-Weibull model and Eyring-lognormal model are all in this category. Another widely studied ALT model is the proportion hazard (PH) model proposed by Cox.[2] Unlike the AFT models, the PH model assumes that the accelerating variables affect the failure rate multiplicatively. In the literature, this model has been extended by Ciampi and Etezadi[3] and Elsayed *et al.*[4] to include many existing ALT models as special cases. More recently, Elsayed and Zhang[5] propose to apply proportional odds model to analyze ALT data. For more comprehensive literature reviews on ALT models, readers are referred to Nelson,[6] Elsayed,[1] Meeker and Escobar,[7] and Escobar and Meeker.[8]

In addition to developing ALT models, extensive research has been conducted on the optimal design of ALT plans following the work by Chernoff.[9] Most traditional ALT plans are formulated for constant-stress loadings. The design objectives usually considered are to minimize the variance of estimated failure time percentile,[13] the determinant of variance-covariance matrix[7,11] (D-optimality) and the asymptotic variance of estimated failure rate.[5] Examples for constant-stress ALT plans can be found in Nelson and Kielpinski,[10] Maxim *et al.*,[11] Meeker and Hahn,[12] and Nelson and Meeker.[13] Regarding ALT using time-varying stress loadings, step-stress ALT plans have been the main focus. Miller and Nelson[14] present the cumulative exposure model and obtain the optimal test plans that minimize the asymptotic variance of maximum likelihood estimate (MLE) of the product's mean life at the normal operating conditions. Bai and Chun[15] obtain the optimal plan for a simple step-stress ALT with competing causes of failure. Chung and Bai[16] consider the optimal design of step-stress ALT using the cumulative exposure model. Khamis and Higgins[17] present 3-step step-stress ALT plans assuming a linear or quadratic life-stress relationship. Xiong[18] investigates statistical inference on the parameters of a simple step-stress ALT model with Type II censoring. Xiong and Milliken[19] study the statistical models in step-stress ALT when the stress change times are random. Xiong and Ji[20] study the optimal design of simple step-stress ALT plan involving grouped and censored data. Recently, Xu and Fei[21] consider the optimal step-stress ALT plans involving two stress variables. For a more comprehensive literature review on ALT plans, readers are referred to Nelson.[22]

1.2. *Motivation and problem description*

In practice, many ALT plans involving different stress loadings can be selected before conducting ALT; however, for most applications, constant- or step-stress

loadings are usually considered. In the literature, most research has been focused on the optimal design of ALT plans when the stress loadings are specified. Although these ALT plans may have desired statistical properties, some of them can not be implemented by a practitioner due to the limited capacity of his/her testing facility (e.g., instantaneous changes of stress levels in step-stress ALT may not be actually implemented due to the limited ramp rate of an environmental test chamber) or the limited quantity of equipment for simultaneous testing. As a result, it is required to design an equivalent test plan using an applicable stress loading. The design problem becomes more attractive if the resulting plan is expected to have more desirable statistical properties in addition to the desired statistical properties. For example, in addition to achieving the low variance for an estimated reliability, the total uncertainty in the unknown parameter estimates measured by the determinant of the variance-covariance matrix is also expected. Furthermore, the equivalent ALT plans can provide more flexible experimental choices in balancing desired statistical properties with other requirements such as testing time and testing units. Hence, studying the equivalency of different ALT plans considering multiple objectives becomes an interesting and challenging problem. However, this problem has not been investigated.

To our best knowledge, the only multiobjective ALT plan in the literature is reported by Tang and Xu.[23] In this multiobjective ALT plan, two conflicting objectives, i.e., the desired level of statistical precision for an estimate of interest and the cost for conducting the test, are considered. The problem is formulated using a simple weighted sum approach. Likewise, the equivalency of different ALT plans under the multiobjective consideration can be addressed either by combining multiple objectives into one composite objective through a utility function or identifying a Pareto optimum solution set. When the first method is applied, only one single final solution/design will be obtained which can be regarded as equivalent to the baseline ALT plan. On the other hand, multiple objective evolutionary algorithms such as NSGA-II can be applied in searching the Pareto optimum solution set. These solutions can be considered to be equivalent to the baseline ALT plan in terms of the multiple desired objectives. However, because the possible huge number of Pareto optimal solutions may be prohibitive for a practitioner to make choices, the Pareto set needs to be pruned before implementation. Many data mining methods such as statistical classification methods may be utilized to provide useful tradeoff information about these alternative solutions. However, such classification methods does not reduce the size of the Pareto optimal set. To facilitate the implementation in ALT, the Pareto optimal solutions need to be further pruned using appropriate solution selection methods. Since choosing one Pareto optimal solution among the whole Pareto set can be considered as a multiobjective selection optimization problem, specific methods such as the Data Envelopment Analysis (DEA) can be applied. This paper is focused on developing a systematic approach to classify the Pareto optimal solutions and reduce the size of the solution set for the design of multiobjective equivalent ALT plans.

The remainder of the paper is organized as follows. Preliminaries regarding ALT plans under Type-I censoring are described in Sec. 2. Section 3 formulates the design of equivalent ALT plans considering multiple objectives and introduces several available solution methods. Section 4 introduces the Pareto solution classification and reduction method for planning multiobjective equivalent ALTs. A numerical example is provided in Sec. 5. Finally, Sec. 6 draws conclusions.

2. ALT Model and Statistical Inference

Throughout this paper, the following assumptions are made.

Assumptions:

1. The stress level S_0 under the product's normal operating condition is constant.
2. The failure time distribution of the product can be described by a log-location-scale distribution as:
$$F(t;\mu,\sigma) = \Phi\left(\frac{\log(t)-\mu}{\sigma}\right), \tag{1}$$
where σ is the scale parameter, μ is the location parameter, and $\Phi(\cdot)$ is the cumulative distribution function (CDF) of a standard location-scale distribution, for which $\mu = 0$ and $\sigma = 1$. Thus, the associated probability density function (*pdf*) is $f(t;\mu,\sigma) = \frac{1}{\sigma t}\phi(\frac{\log(t)-\mu}{\sigma})$, where $\phi(y) = d\Phi(y)/dy$.
3. The scale parameter σ is constant within the range of the stress considered; the location parameter μ depends on the stress level, which can be described by a life-stress relationship: $\mu = g(S;\underline{\beta})$ for a stress level S, where $\underline{\beta}$ is a vector containing the parameters in the relationship.
4. Test units in ALT are independent and subject to singly Type I censoring with the censoring time of t_c.

For convenience, the stress values can be standardized using a normalization scheme. In this paper, the linear normalization method is applied. Let S_H be the highest stress level that can be used in the ALT experiment without introducing different failure mechanisms. Then, accelerated stress level S_i can be standardized as:
$$Z_j = (S_j - S_0)/(S_H - S_0), \quad j = 1,\ldots,k, \tag{2}$$
where $0 < Z_j < 1$. Moreover, the use condition S_0 becomes $Z_0 = 0$ and S_H becomes $Z_H = 1$. Let $\underline{\Theta}$ be the vector containing σ and other parameters in the life-stress relationship after the standardization procedure. This parameter vector will be used throughout the rest of this paper.

2.1. *Likelihood function of ALT data*

Let t_{ji} be the failure time of unit i that fails under stress level Z_j. Under Type I censoring, the log-likelihood function of the ALT data obtained under a specified

stress loading can be generally expressed as:

$$L(\text{Data}|\underline{\Theta}) = \sum_{i=1}^{N}\sum_{j=1}^{k}[\delta_{ji}\log(f_j(t_{ji};\underline{\Theta})) + (1-\delta_{ji})\log(1-F_j(t_{ji};\underline{\Theta}))], \quad (3)$$

where $f_j(\cdot)$ and $F_j(\cdot)$ are the *pdf* and CDF of observed failure times under Z_j, respectively, N is the total number of test units, and

$$\delta_{ji} = \begin{cases} 1 & \text{if unit } i \text{ fails at } Z_j, \\ 0 & \text{otherwise.} \end{cases}$$

For example, for a k-level constant-stress ALT with the single censoring time t_c for all the levels, the log-likelihood function becomes:

$$L(\text{Data}|\underline{\Theta}) = \sum_{j=1}^{k}\sum_{i=1}^{n_j}[\delta_{ji}\log(f_j(t_{ji};\underline{\Theta})) + (1-\delta_{ji})\log(1-F_j(t_{ji};\underline{\Theta}))], \quad (4)$$

where n_j is the number of units allocated to Z_j, and $\sum_{j=1}^{k} n_j = N$.

2.2. *Information matrix and asymptotic variance of MLE*

Throughout the rest of development, it is assumed that the moderate regularity conditions are satisfied. Then, the information matrix associated with an ALT plan can be expressed as:

$$\mathbb{I}_{\underline{\Theta}} = -E\left[\frac{\partial^2 L}{\partial\underline{\Theta}\partial\underline{\Theta}^T}\right]. \quad (5)$$

For instance, for a k-level constant-stress ALT, the information matrix can be expressed as:

$$\mathbb{I}_{\underline{\Theta}} = \sum_{j=1}^{k} n_j I_j = \sum_{j=1}^{k} -n_j\left[\int_0^{t_c}\frac{\partial^2\log(f_j(t;\underline{\Theta}))}{\partial\underline{\Theta}\partial\underline{\Theta}^T}f_j(t;\underline{\Theta})dt + [1-F_j(t_c;\underline{\Theta})]\frac{\partial^2\log(1-F_j(t_c;\underline{\Theta}))}{\partial\underline{\Theta}\partial\underline{\Theta}^T}\right]. \quad (6)$$

where I_j is the information matrix of an observation obtained under Z_j.

Moreover, the asymptotic distribution of the MLE $\underline{\hat{\Theta}}$ follows the multivariate normal distribution with *pdf* of $MVN(\Theta, \sum_{\underline{\hat{\Theta}}})$, where the variance-covariance matrix $\sum_{\underline{\hat{\Theta}}}$ can be estimated by $\mathbb{I}^{-1}_{\underline{\hat{\Theta}}}$ evaluated at the MLE $\underline{\hat{\Theta}}$. This is called the MLE of $\sum_{\underline{\hat{\Theta}}}$ denoted by $\hat{\sum}_{\underline{\hat{\Theta}}}$. Let $G(\underline{\hat{\Theta}})$ be a function of the parameter estimates such as the estimate of reliability function of units under the normal operating condition Z_0 at time T_0: $G(\underline{\hat{\Theta}}) = \overline{F_0}(T_0; Z_0, \underline{\hat{\Theta}})$. Using the delta method (Meeker and Escobar[7]), the asymptotic variance of $G(\underline{\hat{\Theta}})$, denoted by $\text{Asvar}(G(\underline{\hat{\Theta}}))$, can be

approximated by:

$$\text{Asvar}(G(\underline{\hat{\Theta}})) = \left[\frac{\partial G(\underline{\hat{\Theta}})}{\partial \underline{\hat{\Theta}}}\right]^T \hat{\sum}_{\underline{\hat{\Theta}}} \left[\frac{\partial G(\underline{\hat{\Theta}})}{\partial \underline{\hat{\Theta}}}\right]. \tag{7}$$

For notation simplicity, we denote $\underline{c} = \frac{\partial G(\underline{\hat{\Theta}})}{\partial \underline{\hat{\Theta}}}$, which will be used throughout the rest of this paper.

Based on some preliminary ALT, the initial estimate $\underline{\hat{\Theta}}$ can be obtained. Then, the vector of $\underline{c} = \frac{\partial G(\underline{\hat{\Theta}})}{\partial \underline{\hat{\Theta}}}$ will also be known. The ALT planning methodology based on the initial parameter estimates is called local optimal design.[6] As a result, $\hat{\sum}_{\underline{\hat{\Theta}}}$ depends only on those decision variables (e.g., stress levels, censoring time and quantities of testing units) in planning ALT. In other words, an ALT planning process can be regarded as a matrix design problem, where the elements in $\hat{\sum}_{\underline{\hat{\Theta}}}$ are functions of the decision variables. Moreover, the equivalent ALT planning process turns to be a matrix design problem using different stress loadings and different sets of experimental decision variables.

3. Formulation for Multiobjective Design of Equivalent ALT Plans

Before we proceed, several definitions of multiobjective equivalent ALT plans are given first.

Definition 1. (Strict Multiobjective Equivalent ALT Plans). Two ALT plans are strictly equivalent if they generate the same values of all the desired objective functions.

Let $\underline{x}^{(0)}$ and $\underline{x}^{(1)}$ be the vectors containing the decision variables of the baseline ALT plan and the desired equivalent ALT plan, respectively. The number of decision variables in $\underline{x}^{(0)}$ and $\underline{x}^{(1)}$ may be different. Moreover, let $\mathbb{X}|_{\underline{x}^{(0)}}$ and $\mathbb{Y}|_{\underline{x}^{(1)}}$ be the variance-covariance matrices of the unknown parameters of the baseline ALT plan and the equivalent ALT plan, respectively. Then, Definition 1 can be expressed mathematically as:

$$\underline{c}_i^T \mathbb{X}|_{\underline{x}^{(0)}} \underline{c}_i = \underline{c}_i^T \mathbb{Y}|_{\underline{x}^{(1)}} \underline{c}_i, \quad i = 1, 2, \ldots \tag{8}$$

$$H_j(\underline{x}^{(0)}) = H_j(\underline{x}^{(1)}), \quad j = 1, 2, \ldots \tag{9}$$

where each $\underline{c}_i$ is the gradient vector of a specific functional form of $G(\underline{\hat{\Theta}})$ as shown in Eq. (7). Thus, $\underline{c}_i^T \mathbb{X}|_{\underline{x}^{(0)}} \underline{c}_i$ is the target objective related to the estimation accuracy such as the asymptotic variance of reliability estimate at a given mission time under normal operating conditions. Each of the objective functions H_j is called exogenous objective, such as total testing time or cost. For two ALT plans to be strictly equivalent, all the above objectives must be equal for the two ALT plans. However, this requirement is difficult to achieve in most applications. To relax the requirement, the following two alternative definitions would be more practical.

Definition 2. (Utility Value based Equivalent ALT Plans). Two ALT plans are equivalent if the utility values of the objectives in the two ALT plans are equal.

Definition 2 relaxes the strict requirement on the equality of objective values of two ALT plans. In practice, an equal utility value is much easier to achieve. To obtain such multiobjective equivalent ALT plans, a utility function that accurately reflects the decision makers' preferences has to be identified. This becomes the most important task for this method; however in many cases it is difficult to justify the accuracy of the identified utility function.

Definition 3. (Pareto Optimality based Equivalent ALT Plans). Two ALT plans are equivalent if the objective values in the two ALT plans are non-dominated to each other.

Pareto dominance and non-dominance can be determined through multiple pair-wise vector comparison. More specifically, let $\underline{x} = (x_1, x_2, \ldots, x_m)$ and $\underline{y} = (y_1, y_2, \ldots, y_m)$ be two vectors containing the m decision variables. In a minimization problem with l objectives: $f_i(\underline{z}), i \in \{1, 2, \ldots, l\}$, it is said that solution $\underline{x}$ dominates solution $\underline{y}$ if and only if:

$$f_i(\underline{x}) \leq f_i(\underline{y}), \quad \text{for all } i; \quad \text{and} \quad f_i(\underline{x}) < f_i(\underline{y}) \quad \text{for at least one } i \in \{1, 2, \ldots, l\}.$$

In other words, $\underline{x}$ is non-dominated if there is no other $\underline{y} \neq \underline{x}$ such that $f(\underline{y}) \leq f(\underline{x})$.

Compared with the utility value based equivalent ALT plan, the Pareto optimality based equivalent ALT plan is more general. This definition is based on non-dominance concept[24]). It is easier to obtain such equivalent ALT plans and many Pareto optimal equivalent ALT plans can be presented to decision makers. Moreover, it avoids obtaining an inappropriate single "optimal" design due to the misspecification of utility function.

In this paper, both the utility value based equivalent ALT plan and the Pareto optimality based equivalent ALT plan are investigated with the emphasis on the second alternative due to its broader applicability.

3.1. *Formulation of a multiobjective optimization problem*

Let $\underline{x} = [x_1, x_2, \ldots, x_m]$ be a vector containing m decision variables. Mathematically, an optimization problem with l objective functions, P inequality constraints and Q equality constraints can be expressed as:

$$\min_{\underline{x}} \quad f_i(\underline{x}), \quad i = 1, 2, \ldots, l. \tag{10}$$

$$s.t. \quad g_p(\underline{x}) \leq 0, \quad \text{for } p = 1, 2, \ldots, P, \tag{11}$$

$$g_q(\underline{x}) = 0, \quad \text{for } q = 1, 2, \ldots, Q. \tag{12}$$

A variety of approaches can be used to solve this problem. One popular approach is to combine these objectives into one single composite objective so that traditional mathematical programming methods can be applied. To this end, a utility function

needs to be identified according to the preference of one or multiple decision makers. The simplest method is to assume independent preferences among these objectives and apply an additive utility function (e.g., see Tang and Xu[23]). On the other hand, instead of transforming the original problem into a single objective one, the Pareto optimum concept based on non-dominance can be utilized. Under Pareto optimum, a solution set containing all the non-dominated solutions is called the Pareto optimal set or Pareto front of the multiobjective optimization problem. By introducing the Pareto optimum concept, more choices may be presented to different decision makers with different perspectives.

In many cases, however, the number of solutions in a Pareto optimal solution set is too large, so a Pareto solutions pruning method is needed to assist decision makers in determining several representative solutions. The proposed method in this paper incorporates SOM and DEA, which not only provides good tradeoff information among different Pareto optimal solutions but also reduces the number of solutions to a workable size. It bridges the gap between single solution and Pareto optimal set. The detailed methodology will be elaborated in Section 4.

3.2. *Formulation of multiobjective equivalent ALT plans*

For a baseline ALT plan, the associated decision variables $\underline{x}^{(0)}$ are known, then the variance-covariance matrix $\mathbb{X}|_{\underline{x}^{(0)}}$ as well as all the objectives can be determined. To develop multiobjective equivalent ALT plans, the following widely used objectives can be considered. The first objective is the total testing time (Type I censoring time) t_c. The second one is the accuracy of parameter estimates, which can be measured by the determinant of the variance-covariance matrix of the equivalent ALT plan, i.e., $\det(\mathbb{Y})$. The third one is to achieve the desired accuracy of a reliability estimate same as (or close to) the baseline ALT plan. This objective can be measured by the deviation of the objective value obtained in the equivalent ALT plan from the one in the baseline ALT plan. For simplicity, all the objectives are included in a minimization problem given by:

$$f_1(\underline{x}^{(1)}) = t_c, \quad f_2(\underline{x}^{(1)}) = \det(\mathbb{Y}|_{\underline{x}^{(1)}}), \quad f_3(\underline{x}^{(1)}) = \|\underline{c}^T\mathbb{X}|_{\underline{x}^{(0)}}\underline{c} - \underline{c}^T\mathbb{Y}|_{\underline{x}^{(1)}}\underline{c}\|, \quad (13)$$

where $\underline{x}^{(1)}$ contains the decision variables of the desired equivalent ALT plan, such as the stress levels, the unit allocation at each stress level and the censoring time of the test. The constraints to be included are usually imposed by the limited testing resources such as the allowed total testing time and total number of testing units.

To formulate the design of multiobjective equivalent ALT plans, both the utility function method and the Pareto optimum method can be utilized.

Utility value based equivalent ALT plans

Suppose the utility function can be identified as: $U(f_1(\underline{x}^{(1)}), f_2(\underline{x}^{(1)}), f_3(\underline{x}^{(1)}))$. For example, under preference independence and additive utility assumptions, one may

obtain a utility function as weighted sum of the multiple objectives, i.e.:

$$U(f_1(\underline{x}^{(1)}), f_2(\underline{x}^{(1)}), f_3(\underline{x}^{(1)})) = w_1 f_1(\underline{x}^{(1)}) + w_2 f_2(\underline{x}^{(1)}) + w_3 f_3(\underline{x}^{(1)}). \quad (14)$$

Then the traditional single objective optimization methods can be applied to solve the above integrated single objective optimization problem with Eq. (14) being the objective function. Usually, the resulting equivalent ALT plan is prone to change as the utility function is varied.

Pareto optimality based equivalent ALT plans

For a Pareto optimal solution, improving one objective would sacrifice at least one other objective. One method to obtain the Pareto optimal solutions is to utilize a utility function identified. Each time a set of weights (sum to one) are generated from the uniform distribution, one Pareto optimal solution can be obtained, if exists. By repeating this procedure, the desired number of Pareto optimal solutions will be obtained. This method is computationally inefficient and can not guarantee a good variety of solutions. To achieve a good representations of the Pareto solution space, a genetic algorithm based searching procedure will be utilized in this paper, which is more computationally efficient than the random weights generating approach. The resulting Pareto optimality based equivalent ALT plans are non-dominated in terms of the multiple objectives: $f_i(\underline{x}^{(1)})$, $i = 1, 2, 3$.

3.3. *Solution techniques*

The utility function based equivalent ALT plans are relatively easy to obtain. However, because the utility function is sensitive to different decision makers and also sensitive to the acquired tradeoff information during interaction with decision makers, it is attractive to obtain Pareto optimality based equivalent ALT plans. In this section, we focus on searching the Pareto optimality based equivalent ALT plans.

The traditional genetic algorithm (GA) was developed by Holland,[25] which is a particular class of evolutionary algorithms. It starts with a population of random individuals (called chromosomes). The crossover and mutation operators are used to generate new solutions at each generation. For each generation, each solution is evaluated in terms of a fitness function, and individuals with higher fitness values are ranked at the top while individuals with low-fitness values are likely to be eliminated from the current population. The algorithm continues for a pre-determined number of generations or until no additional improvement is observed.

To solve a multi-objective optimization problem, the following multiobjective GAs, referred to as MOEA, have been developed:

- Vector evaluated GA by Shaffer[26];
- MOGA by Fonseca and Flemming[27];
- Niched-Pareto GA by Horn *et al.*[28];

- non-dominated genetic algorithm (NSGA) developed by Srinivas and Deb[29];
- Strength Pareto evolutionary algorithm by Zitzler and Thiele[30];
- NSGA-II by Deb *et al.*[31,32,33]
- MOMS-GA by Taboada *et al.*[34]

Specially, NSGA uses a non-dominated sorting procedure.[29] It applies a ranking method that emphasizes those good solutions and tries to maintain them in the population. Through a sharing method, this algorithm maintains the diversity in the population. The algorithm explores different regions in the Pareto front. This algorithm can accommodate many objectives and constraints and is very efficient in obtaining good Pareto optimal sets. As an improved version of NSGA, NSGA-II utilizes a fast non-dominated sorting genetic algorithm. This method is more computationally efficient, non-elitism preventing, and less dependent on sharing parameter for diversity preservation. In this paper, the multiobjective equivalent ALT plans are obtained using NSGA-II with moderate modification to effectively search the Pareto front.

4. Classification and Pruning of Pareto Solutions

4.1. *solution classification using self-organizing map (SOM)*

For the design of multiobjective equivalent ALT plans, the size of Pareto optimal solution set may be extremely large. Therefore, it is very useful to classify the Pareto optimal solutions first in order to better understand the performance characteristics of these solutions and to do meaningful tradeoff. Moreover, such solution identification processes can maintain the completeness of the Pareto optimal solution set.

There are many statistical classification methods, which have been widely used in mining useful information from huge data sets. Based on the amount of prior knowledge about the original data, either unsupervised statistical classification methods (e.g., k-means and SOM) or supervised statistical classification alternatives (e.g., artificial neural network and support vector machine) may be utilized. Specially, when no or very few prior information about the data is available, the unsupervised classification approach is more appropriate in classifying the data.

For the design of multiobjective equivalent ALT plans, each Pareto optimal solution can be treated as an input vector containing the determinant of the variance-covariance matrix of the equivalent ALT plan, the testing termination time, and the deviation from the target accuracy of reliability estimate of the baseline ALT plan. Because there is no prior information about the cluster to which an input vector belongs, the unsupervised classification method would be more appropriate. Specifically, SOM is used to cluster Pareto optimal sets based on performance similarities.

SOM is a special artificial neural network with a single layer feedforward structure.[35] It generates a set of representations (usually two-dimensional or three-dimensional) for multi-dimensional input vectors while preserving the topological

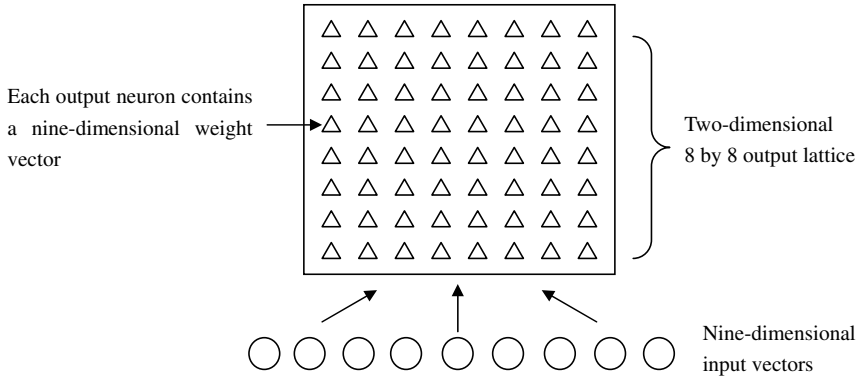


Fig. 1. SOM single layer feedforward network.

properties measured by the similarity of these input vectors, such as the same or similar distances and angles between them. More specifically, each input vector is connected to all output neurons, and a weight vector with the same dimensions as the input vectors is attached to each neuron (see Fig. 1 for the case with a two-dimensional output lattice). Usually, the number of dimensions of an input vector is much higher than that of the output lattice, so the mapping from the input space to the output space can be regarded as a dimension reduction process. In fact, it is very useful to take the advantage of the reduced dimension for understanding the performance of the alternative solutions.

During the SOM's training process, a competitive learning technique is utilized. When a training sample (input vector) is given to the network, its Euclidean distance to all weight vectors is computed. The neuron with the weight vector that is most similar to the input is called the Best Matching Unit (BMU). The weights of the BMU and neurons close to it in the SOM lattice are then adjusted towards the input vectors. The magnitude of the adjustment decreases with time and is smaller for those neurons that are far away from the BMU in the lattice. The weight $\underline{w}(t)$ is updated iteratively as:

$$\underline{w}(t+1) = \underline{w}(t) + \rho(v,t)\alpha(t)[\underline{\xi}(t) - \underline{w}(t)], \tag{15}$$

where $\underline{w}(t)$ and $\underline{w}(t+1)$ are the weight vectors at step t and step $t+1$, respectively; $\underline{\xi}(t)$ is the input vector; $\alpha(t)$ is the learning coefficient that is monotonically decreasing with time; $\rho(v,t)$ is the neighborhood function, which has a smaller value when the neuron is far away from the BMU (as determined by v, the Euclidean distance) and decreases with time. For instance, the Gaussian neighborhood function given by: $\rho(v,t) = \exp(-v^2/\sigma^2)$, has been widely used, where σ^2 is the width parameter that gradually decreases over time. This updating process can be performed for a given number of iterations or until $\underline{\xi}(t)$ approaches the weight vector $\underline{w}(t)$. To ensure the quality of this training process, the representatives of all the possible

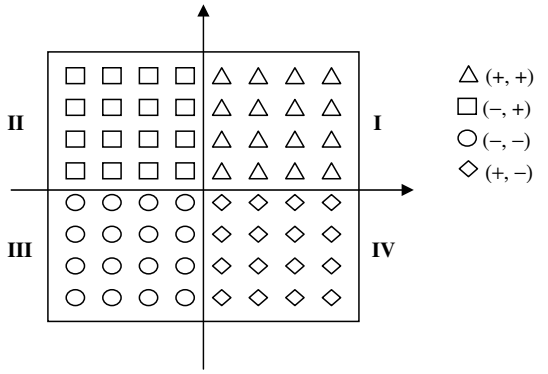


Fig. 2. Two-dimensional output lattice representation in an output lattice.

input vectors need to be selected as the training samples. Eventually, output nodes are associated with groups or patterns corresponding to the input vectors. In the subsequent mapping process, a new input vector is mapped to a specific location on the lattice based on its similarity to the weight vector of a specific neuron.

Figure 2 shows an example two-dimensional output lattice after training. The coordinates provide a visual representation of the input vectors in the output space. During the mapping process, those input vectors that are similar to the weight vectors of the neurons in quadrant I are assigned coordinates with signs of $(+, +)$ in the output lattice. Similarly, the input vectors having a good similarity with the weight vectors of the neurons in quadrant II are allocated coordinates with signs of $(-, +)$, and so on for quadrant III and quadrant IV. Unlike the k-means method that only minimizes the mean squared error in the Euclidian distance among the input vectors, SOM measures the similarity of input vectors by the Euclidian distance as well as the angle between them by updating the weight vectors iteratively.[36] Such training process results in the topological preservation from the input vectors to the output lattice map. Because of these advantages, SOM is utilized to classify the Pareto optimal solutions after the NSGA-II algorithm in searching the equivalent ALT plans.

4.2. *Solution reduction using data envelopment analysis (DEA)*

Even though the classification results are informative, the number of solutions in each cluster may still be prohibitive for a decision maker to make choices. At this point, selecting representative solutions from each cluster itself can be regarded as a multiobjective optimization problem, also called multiple objective selection optimization (MOSO) problem.[37] If appropriately applied, an MOSO method can significantly reduce the size of each cluster of Pareto optimal solutions.

A special MOSO method is the data envelopment analysis (DEA) method. It is a linear programming based technique for measuring and comparing the relative

performance of decision making units (DMUs) with multiple inputs (e.g., cost type criteria) and outputs (e.g., benefit type criteria).[38] To implement the DEA method in MOSO, each alternative solution is treated as a DMU, and all the DMUs are assumed to be homogeneously comparable. The DEA relative efficiency of a solution is defined as the ratio between the weighted outputs and the weighted inputs of the solution. By comparison, those solutions with lower relative efficiency values can be eliminated.

Considering a problem involving l DMUs, each of which has m inputs and n outputs, then the relative efficiency (RE) of the p^{th} DMU can be expressed as:

$$RE_p = \frac{\sum_{j=1}^{n} u_j y_{jp}}{\sum_{i=1}^{m} v_i x_{ip}}, \quad p = 1, 2, \ldots, l; \tag{16}$$

$$u_j, v_i \geq \epsilon, \quad i = 1, 2, \ldots, m; \quad j = 1, 2, \ldots, n, \tag{17}$$

where u_j and v_i are the weights for the outputs and inputs, respectively, and ϵ is a small positive quantity which guarantees the weights are nonnegative. Since different DMUs may utilize different strategies to achieve their highest relative efficiency values, a specific weight set for each DMU is more practical instead of pursuing a common set of weights for all DMUs.[39] Consequently, the relative efficiency of a specific DMU p_0 can be obtained as a solution to the following problem:

$$\max_{u_j, v_i, \forall i,j} \quad RE_{p_0} = \frac{\sum_{j=1}^{n} u_j y_{jp_0}}{\sum_{i=1}^{m} v_i x_{ip_0}} \tag{18}$$

$$s.t. \quad \frac{\sum_{j=1}^{n} u_j y_{jp}}{\sum_{i=1}^{m} v_i x_{ip}} \leq 1, \quad p = 1, 2, \ldots, l, \tag{19}$$

$$u_j, v_i \geq \epsilon, \quad i = 1, 2, \ldots, m; \quad j = 1, 2, \ldots, n, \tag{20}$$

where ϵ is a small positive quantity. The decision variables of the problem are those weights, and the solution contains a weight set most favorable to the unit and the value of its relative efficiency. Moreover, maximizing a fraction or ratio depends on the relative magnitude of the numerator and denominator but not on their individual values. Therefore, the same result can be obtained by setting the denominator equal to a constant and maximizing the numerator instead. As a result, the above fractional linear programming problem can be transformed to a general

linear programming problem as:

$$\max_{u_j, v_i, \forall i,j} \quad RE_{p_0} = \sum_{j=1}^{n} u_j y_{jp_0} \tag{21}$$

$$s.t. \quad \sum_{i=1}^{m} v_i x_{ip_0} = 1, (\text{normalization}) \tag{22}$$

$$\sum_{j=1}^{n} u_j y_{jp} - \sum_{i=1}^{m} v_i x_{ip} \leq 0, \quad p = 1, 2, \ldots, l, \tag{23}$$

$$u_j, v_i \geq \epsilon, \quad i = 1, 2, \ldots, m; \quad j = 1, 2, \ldots, n. \tag{24}$$

To obtain the efficiencies of the entire set of units, it is necessary to solve a linear program for each unit. Clearly, as the objective function varies from one problem to another, the weights obtained for each unit may be different. Moreover, when applying DEA, all DMUs attempt to select their most favorable weights; therefore there may be more than one efficient unit whose relative efficiency is equal to one. When there are two inputs or two cost type criteria, all the DMUs with the relative efficiency values equal to one provide an efficient frontier in a two-dimensional input space. For example, A, B and C in Fig. 3 are efficient whereas D and E are not. Therefore, the efficient frontier goes through A, B and C. If there are three inputs, an envelopment surface may be formed by connecting those points whose relative efficiency values are equal to one.

In the context of the design of multiobjective equivalent ALT plans, those ALT plans with high efficiency values are preferred and will be selected from the clusters of Pareto solutions. To formulate the MOSO for the design of multiobjective equivalent ALT plans, all the Pareto optimal solutions in each cluster is considered as DMUs. Since all the three objectives in each equivalent ALT plan are of the minimization type, these objectives are all treated as inputs to the DEA model. A common single constant output is utilized as a dummy variable, and different choices for the constant value do not affect the relative efficiency value of the DMUs.

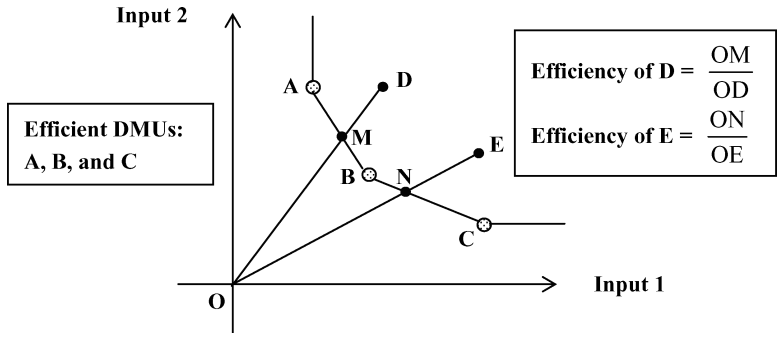


Fig. 3. Illustration of the DEA efficiency with two inputs.

A higher relative efficiency value indicates that a smaller input (e.g., total testing time) is consumed while the greater or at least the same amount of output is produced. Often, those solutions with high relative efficiencies (equal to one) are preferred, and others can be eliminated from the cluster. This is a strong statement because if even the favorable weight set can not achieve the relative efficiency value of one, that solution must be an inefficient solution. This method is appropriate in particular when decision makers have not provided any preferences to those objective functions in our ALT planning problem. An alternative method to prune the Pareto optimal set is based on an ordinal ranking of objective functions, as described by Taboada *et al.*[39]

5. Numerical Example

For demonstration, a simple log-linear exponential ALT model is considered. In this model, the failure rate λ depends on stress level Z as: $\lambda(Z) = \exp(\theta_0 + \theta_1 Z)$, so:

$$F(t, Z, \underline{\Theta}) = 1 - \exp(-\exp(\theta_0 + \theta_1 Z)t), \tag{25}$$

where $\underline{\Theta} = [\theta_0, \theta_1]$. Suppose the baseline estimates of the parameters are: $\hat{\theta_0} = -7.3646$ and $\hat{\theta_1} = 0.3398$. Of our interest is the estimate of reliability function $\overline{F}(t, Z, \underline{\Theta})$ of the product under the standardized use condition $Z_0 = 0$ at the specific mission time $T_0 = 1000$ hours. The partial derivatives of the reliability function with respect to the model parameters are:

$$\frac{\partial \overline{F}(T_0, Z_0, \underline{\Theta})}{\partial \theta_0} = -T_0 \exp(\theta_0) \exp(-T_0 \exp(\theta_0)), \tag{26}$$

$$\frac{\partial \overline{F}(T_0, Z_0, \underline{\Theta})}{\partial \theta_1} = 0. \tag{27}$$

So, the gradient vector is $\underline{c} = \left[\frac{\partial \overline{F}(T_0, Z_0, \hat{\underline{\Theta}})}{\partial \hat{\theta}_0}, 0\right] = [-0.33618, 0]$.

Suppose that the baseline ALT plan is a three-level constant-stress run-to-failure test plan with stress levels: $Z_1 = 0.5$, $Z_2 = 0.75$ and $Z_3 = 1$, and ten units are allocated to each stress level. The empirical variance-covariance matrix of $\hat{\underline{\Theta}}$ obtained from a pilot test is given by:

$$\mathbb{X} = \begin{pmatrix} 0.4836 & -0.5895 \\ -0.5895 & 1.1711 \end{pmatrix}.$$

Our goal is to find a three-level constant-stress ALT plan that is equivalent to the baseline ALT plan considering multiple objectives. For the desired equivalent plan, the total number of test units is $N = 70$, and Type I censoring is adopted with censoring time $t_c \leq 150$ unit time. Moreover, the high stress level $Z_3 = 1$ and the unit allocation to this stress level $n_3 = 15$ are predetermined; Fig. 4 shows the layout of the desired ALT plan. The decision variables of the desired equivalent ALT plan are $\underline{x}^{(1)} = [Z_1, Z_2, t_c, n_1, n_2]$, where the low stress level Z_1 and the median stress levels Z_2 should satisfy $0 < Z_1 < Z_2 < Z_3 = 1$, and the numbers of units n_1 and n_2 allocated to these two levels should be nonnegative integers.

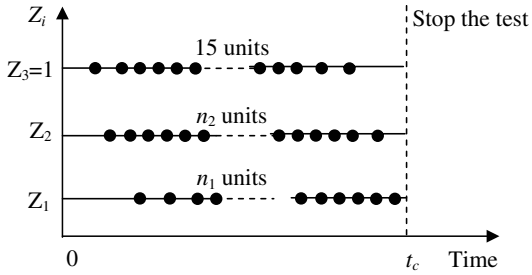


Fig. 4. Design layout of the desired ALT plan.

From Eq. (3), the associated variance-covariance matrix of the desired ALT plan is given by:

$$\mathbb{Y} = \begin{pmatrix} I_{11} & I_{12} \\ I_{12} & I_{22} \end{pmatrix}^{-1},$$

where

$$I_{11} = E\left[-\frac{\partial^2 L}{\partial \theta_0^2}\right] = \sum_{j=1}^{3} n_j[1 - \exp(-\exp(\theta_0 + \theta_1 Z_j)t_c)];$$

$$I_{12} = E\left[-\frac{\partial^2 L}{\partial \theta_0 \theta_1}\right] = \sum_{j=1}^{3} n_j Z_j[1 - \exp(-\exp(\theta_0 + \theta_1 Z_j)t_c)];$$

$$I_{22} = E\left[-\frac{\partial^2 L}{\partial \theta_1^2}\right] = \sum_{j=1}^{3} n_j Z_j^2[1 - \exp(-\exp(\theta_0 + \theta_1 Z_j)t_c)].$$

In this application, the three objectives addressed in Eq. (13) are considered. In addition to achieving the reliability estimate as accurate as the baseline ALT plan, the testing time and accuracy in parameter estimates are also taken into account. It is expected to minimize all the three objectives. Then, the equivalent ALT plan design problem can be expressed as:

$$\min_{\underline{x}^{(1)}} \; f_1(\underline{x}^{(1)}) = t_c\,, f_2(\underline{x}^{(1)}) = \det(\mathbb{Y}|_{\underline{x}^{(1)}})\,,$$

$$f_3(\underline{x}^{(1)}) = \|\underline{c}^T \mathbb{X}_{\underline{x}^{(0)}} \underline{c} - \underline{c}^T \mathbb{Y}|_{\underline{x}^{(1)}} \underline{c}\| \tag{28}$$

$$s.t. \;\; 0 \le Z_1 \le Z_2 \le 1, \tag{29}$$

$$n_l + n_2 = N - n_3 = 55, \tag{30}$$

$$n_1, n_2 \in 1, 2, \ldots, \tag{31}$$

$$t_c \le 150, \tag{32}$$

$$\underline{x}^{(1)} = [Z_1, Z_2, t_c, n_1, n_2]. \tag{33}$$

This multiobjective optimization problem is solved using the NSGA-II algorithm. The algorithm is modified to deal with both continuous and discrete decision variables as shown in Eq. (33). With an initial population of 1000 chromosomes and 50 generations, 43 Pareto optimal solutions are obtained as shown in Table 1, which are also plotted in Fig. 5.

Table 1. Pareto Optimal Solutions for the Equivalent ALT plan.

Solution Index	Z_1	Z_2	t_c	n_1	n_2	$f_1(\underline{x}^{(1)})$	$f_2(\underline{x}^{(1)})$	$f_3(\underline{x}^{(1)})$
Sol #1	**0.0000**	**0.6371**	**126.44207**	**24**	**31**	**126.44207**	**0.1735**	**4.9726E-05**
Sol #2	0.1194	0.8945	129.81916	34	21	129.81916	0.1372	1.0739E-02
Sol #3	**0.0479**	**1.0000**	**142.96019**	**25**	**30**	**142.96019**	**0.0934**	**1.8769E-02**
Sol #4	**0.1118**	**0.7765**	**93.978816**	**45**	**10**	**93.978816**	**0.2862**	**9.1200E-04**
Sol #5	**0.0834**	**0.8785**	**128.15005**	**30**	**25**	**128.15005**	**0.1378**	**1.6742E-02**
Sol #6	0.1592	0.7927	106.75714	45	10	106.75714	0.2451	8.7293E-04
Sol #7	0.1441	0.6115	125.12839	36	19	125.12839	0.2059	1.0129E-02
Sol #8	**0.0236**	**0.9688**	**142.73986**	**25**	**30**	**142.73986**	**0.0943**	**1.4922E-03**
Sol #9	0.1220	0.6013	93.863585	45	10	93.863585	0.3252	2.8946E-03
Sol #10	**0.0781**	**1.0000**	**148.90315**	**27**	**28**	**148.90315**	**0.0885**	**1.0073E-02**
Sol #11	0.1249	0.5540	84.115018	52	3	84.115018	0.4057	3.8704E-03
Sol #12	0.1522	0.7809	106.67297	45	10	106.67297	0.2438	9.6764E-03
Sol #13	0.0000	0.6797	126.6383	24	31	126.6383	0.1669	1.2992E-02
Sol #14	0.0000	0.6236	126.33895	24	31	126.33895	0.1755	4.4287E-03
Sol #15	0.1325	0.8693	129.79183	34	21	129.79183	0.1460	4.7033E-03
Sol #16	0.0040	0.6699	126.30502	24	31	126.30502	0.1706	1.5056E-02
Sol #17	0.0860	0.9532	148.98464	27	28	148.98464	0.0967	7.5736E-04
Sol #18	**0.0837**	**0.1199**	**74.099977**	**19**	**36**	**74.099977**	**0.5014**	**1.1708E-02**
Sol #19	0.1440	0.9169	83.804688	54	1	83.804688	0.4109	8.2141E-03
Sol #20	**0.0671**	**0.0875**	**74.084375**	**19**	**36**	**74.084375**	**0.4773**	**1.8867E-02**
Sol #21	0.0490	0.3743	88.54316	32	23	88.54316	0.3677	8.7809E-03
Sol #22	0.1330	1.0000	83.933335	54	1	83.933335	0.3957	7.7192E-03
Sol #23	0.0446	0.4069	88.566504	32	23	88.566504	0.3644	1.8281E-02
Sol #24	0.0000	0.6658	126.48041	24	31	126.48041	0.1694	9.2928E-03
Sol #25	0.0000	0.6627	126.54233	24	31	126.54233	0.1697	8.0530E-03
Sol #26	0.0000	0.6439	126.51426	24	31	126.51426	0.1724	1.9784E-03
Sol #27	0.0939	0.9799	148.93013	27	28	148.93013	0.0942	5.9388E-03
Sol #28	0.0357	0.3421	88.587327	32	23	88.587327	0.3583	1.6686E-02
Sol #29	0.1448	0.9345	129.81825	34	21	129.81825	0.1367	1.5812E-02
Sol #30	0.1623	0.3697	91.923872	50	5	91.923872	0.3765	5.8377E-03
Sol #31	0.0805	0.9842	149.0567	27	28	149.0567	0.0910	7.5195E-03
Sol #32	**0.0039**	**0.1568**	**74.186467**	**19**	**36**	**74.186467**	**0.4876**	**4.3670E-03**
Sol #33	0.0000	0.6518	126.46881	24	31	126.46881	0.1715	4.7949E-03
Sol #34	0.0000	0.6875	126.45648	24	31	126.45648	0.1662	1.5971E-02
Sol #35	0.0000	0.1391	74.191314	19	36	74.191314	0.4778	9.9078E-03
Sol #36	0.0000	0.6608	126.52541	24	31	126.52541	0.1700	7.5291E-03
Sol #37	**0.0907**	**0.9862**	**149.07692**	**27**	**28**	**149.07692**	**0.0925**	**1.9976E-03**
Sol #38	0.0000	0.6503	126.54044	24	31	126.54044	0.1715	4.0326E-03
Sol #39	**0.0000**	**0.1340**	**74.181361**	**19**	**36**	**74.181361**	**0.4755**	**1.3369E-02**
Sol #40	0.0426	0.9817	142.73271	25	30	142.73271	0.0956	1.5211E-02
Sol #41	0.0000	0.6631	126.5154	24	31	126.5154	0.1697	8.3082E-03
Sol #42	**0.0262**	**0.9524**	**142.73405**	**25**	**30**	**142.73405**	**0.0971**	**6.0611E-04**
Sol #43	**0.0000**	**0.1401**	**74.175598**	**19**	**36**	**74.175598**	**0.4784**	**9.1094E-03**

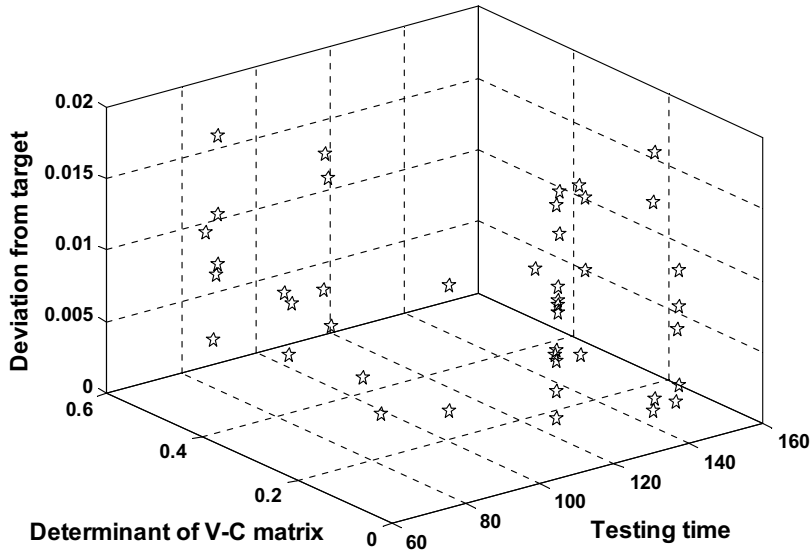


Fig. 5. Original Pareto optimal solutions in a 3-D space.

To obtain an informative tradeoff about these Pareto optimal designs, SOM is applied using Neuralwork pro software to classify these Pareto optimal solutions. A two-dimensional map of the original solutions and the corresponding solutions in the three-dimensional space after the classification are shown in Figs. 6 and 7,

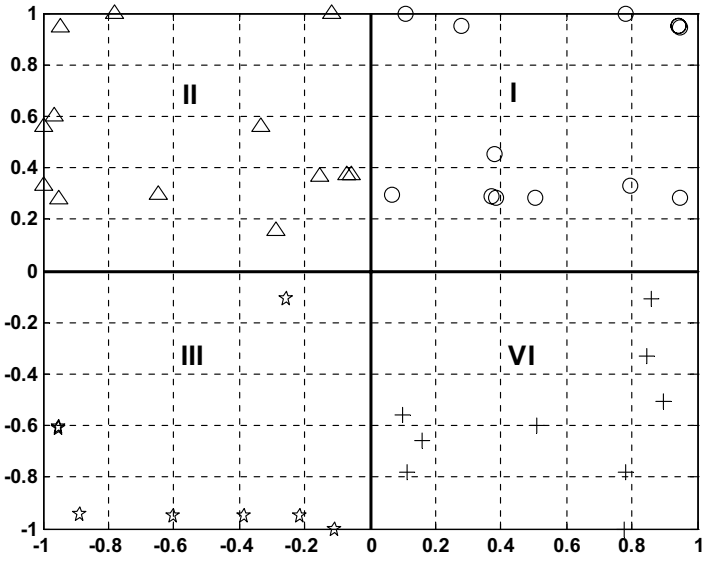


Fig. 6. Pareto optimal solutions map in a 2-D space after SOM classification.

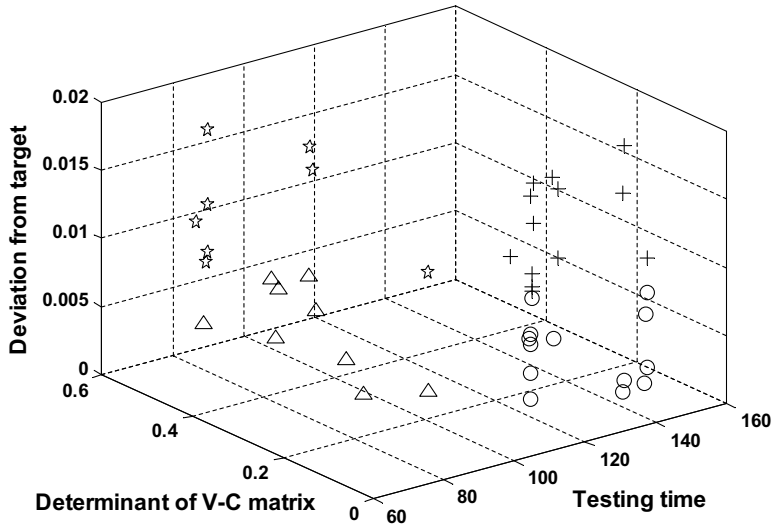


Fig. 7. Pareto optimal solutions in a 3-D space after SOM classification.

respectively. From Fig. 6, one can see that the tradeoff information is very clear for four classes. For example, the designs falling in the quadrant I (with ∘ legend in Fig. 6) have a relative long testing time, small value of the determinant of the variance-covariance matrix, and a small deviation of estimation accuracy from the baseline ALT plan. Similarly, the characteristics of the designs in other three quadrants can also be easily identified. Compared to Fig. 7, the classification results presented in the two-dimensional space is much easier to identify. Such dimension reduction would be more useful when the number of objectives becomes large such that visualization of the high dimensional solutions is very difficult, if not impossible.

Although the SOM classification provides useful tradeoff information for all the Pareto optimal solutions, it is still difficult to choose one design from many alternatives. It is preferred to reduce the number of Pareto optimal solutions to a small set of representative solutions for implementation. For this purpose, the DEA method is applied, where the three objectives of each solution are regarded as the inputs to a DMU and a single common constant (equal to one) is used as the dummy output of all DMUs. For each Pareto optimal design, the relative efficiency needs to be evaluated. For example, the formulation for evaluating the relative efficiency of Sol #1 in Table 1 can be expressed as:

$$\max_{u_1,v_1,v_2,v_3} \quad RE_1 = u_1 \tag{34}$$

$$s.t. \quad \sum_{i=1}^{3} v_i x_{i1} = 1, (\text{normalization}) \tag{35}$$

$$u_1 - \sum_{i=1}^{3} v_i x_{ip} \leq 0, \quad p = 1, 2, \ldots, 43, \tag{36}$$

$$u_1, v_1, v_2, v_3 \geq \epsilon. \tag{37}$$

This linear programming problem intends to search favorable weights for Sol #1 (i.e., DMU1) while satisfying the constraint that each weighted output should be greater than the weighted inputs from an economic perspective. Each of the 43 Pareto solutions will be evaluated using DEA, and those designs with relative efficiency less than 1 will be eliminated from the Pareto solution set. This optimization problem can be solved using either MS Excel or Matlab.

After applying the DEA efficiency evaluation for each class, 13 designs (i.e., Sol #1, Sol #3, etc., which are highlighted in Table 1) remain in the suggested solution set as shown in Fig. 8. These solutions balance different preferences of decision makers and the size of Pareto optimal designs. Since the size of the solution set is small, it would be easy to choose one deign based on a specific need.

It is necessary to mention the difference between the discussion about the Pareto optimality and DEA efficiency[41] and the idea of our application. The former considers the achieved values of those constraints as the inputs and the objective values as the outputs. In our application, however, part (or all) of the objectives are treated as inputs and others as outputs from the economic perspective. More specifically, in this numerical example we treat the three minimization objectives as inputs and a dummy variable is used as an output in the DEA procedure.

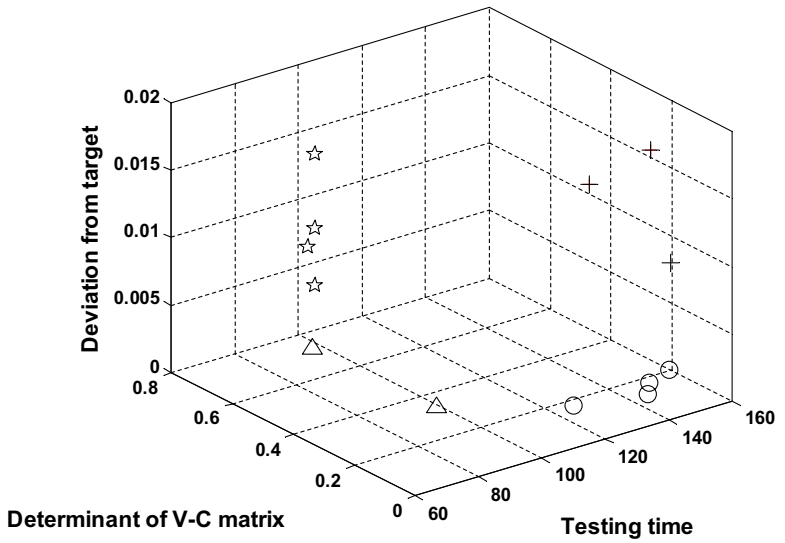


Fig. 8. Reduced Pareto optimal solutions after DEA efficiency evaluation.

6. Conclusions

A multiobjective framework for the design of equivalent ALT plans is investigated in this paper. The multiobjective formulation in conjunction with the proposed Pareto solution classification and reduction technique provide decision makers with more efficient alternative solutions. This systematic solution generating and data mining approach avoids missing the most preferred solutions, and the final decision would be much easier to make as the resulting representative solutions is much smaller in size compared to the original Pareto solution set.

This work considers ALT involving one stress type. However, many products are usually subjected to multiple stresses such as temperature, humidity, electric current, and vibration during operation. To study the reliability of such products, it is necessary to subject test units to multiple stresses in ALT. Since more flexibility and design perspectives may be involved, the design of equivalent ALT plans becomes more challenging when multiple objectives and multiple stresses are considered simultaneously. This problem will be investigated in our future research.

Acknowledgment

The authors would like to thank the editor and referees for their insightful comments that greatly improved the content of this paper. This work was partially supported by the National Science Foundation under grant CMMI-0855812 and the U.S. Nuclear Regulatory Commission under grant NRC-38-08-943.

References

1. E. A. Elsayed, *Reliability Engineering* (Adison Wesley Longman, New York, 1996).
2. D. R. Cox, Regression models and life tables, *J. R. Statist. Soc. B* **34** (1972) 187–220.
3. A. Ciampi and J. Etezadi-Amoli, A general model for testing the proportional hazards and the accelerated failure time hypotheses in the analysis of censored survival data with covariates, *Communications in Statistics-Theory and Methods* **14** (1985) 651–667.
4. E. A. Elsayed, H. T. Liao and X. D. Wang, An extended linear hazard regression model with application to time-dependent-dielectric-breakdown of thermal oxides, *IIE Transactions on Quality and Reliability* **38**(1) (2006) 1–12.
5. E. A. Elsayed and H. Zhang, Optimum multiple-stress-type accelerated life testing plans based on proportional odds model with simple step-stress loading, *Journal Europen des Systmes Automatiss* **40** (2006) 745–762.
6. W. B. Nelson, *Accelerated Testing: Statistical Models, Test Plans, and Data Analysies* (Wiley, New York, 1990).
7. W. Q. Meeker and L. A. Escobar, *Statistical Methods for Reliability Data* (John Wiley & Sons, New York, 1998).
8. L. A. Escobar and W. Q. Meeker, A review of accelerated life test models, *Statistical Sciences* **21** (2006) 552–577.
9. H. Chernoff, Optimal accelerated life design for estimation, *Technometrics* **4** (1962) 381–401.
10. W. B. Nelson and T. J. Kielpinski, Theory for optimum accelerated life tests for normal and lognormal life distributions, *Technometrics* **18** (1976) 105–114.

11. L. D. Maxim, A. D. Hendrickson and D. E. Cullen, Experimental design for sensitivity testing: the Weibull model, *Technometrics* **19** (1977) 405–412.
12. W. Q. Meeker and G. J. Hahn, How to plan an accelerated life test — some practical guidelines, *Statistical Techniques* **10** (1985), ASQC Basic reference in QC.
13. W. B. Nelson and W. Q. Meeker, Theory for optimum accelerated censored life tests for Weibull and extreme value distributions, *Technometrics* **20**(2) (1978) 171–177.
14. R. Miller and W. B. Nelson, Optimum simple step-stress plans for accelerated life testing, *IEEE Transactions on Reliability* **32** (1983) 59–65.
15. D. S. Bai and Y. R. Chun, Optimum simple step-stress accelerated life tests with competing causes of failures, *IEEE Transactions on Reliability* **40** (1991) 622–627.
16. S. W. Chung and D. S. Bai, Optimal designs of simple step-stress accelerated life tests for lognormal lifetime distributions, *International Journal of Reliability, Quality and Safety Engineering* **5** (1998) 315–336.
17. I. H. Khamis and J. J. Higgins, Optimal 3-step step-stress tests, *IEEE Transactions on Reliability* **45** (1996) 341–345.
18. C. Xiong, Inference on a simple step-stress model with type-II censored exponential data, *IEEE Transactions on Reliability* **47** (1998) 142–146.
19. C. Xiong and G. A. Milliken, Step-stress life-testing with random stress-change times for exponential data, *IEEE Transactions on Reliability* **48** (1999) 141–148.
20. C. Xiong and M. Ji, Analysis of grouped and censored data from step-stress life-testing, *IEEE Transactions on Reliability* **53** (2004) 22–28.
21. H. Y. Xu and H. L. Fei, Planning step-stress accelerated life tests with two experimental variables, *IEEE Transactions on Reliability* **56** (2007) 569–579.
22. W. B. Nelson, A bibliography of accelerated test plans, *IEEE Transactions on Reliability* **54** (2005) 194–197.
23. L. C. Tang and K. Xu, A multiple objective framework for planning accelerated life tests, *IEEE Transactions on Reliability* **54** (2005) pp. 58–63.
24. C. W. Kirkwood, *Strategic decision making* (Duxbury Press, 1997).
25. J. Holland, *Adaptation in natural and artificial systems* (U. Michigan Press, 1975).
26. J. D. Schaffer, Multiple objective optimization with vector evaluated genetic algorithms, *Genetic Algorithms and Their Applications: Proceedings of the First International Conference on Genetic Algorithms* (Hillsdale, NJ. 1985).
27. C. M. Fonseca and P. J. Fleming, Genetic algorithms for multiobjective optimization: formulation, discussion and generalization, in *Proc. Fifth Int. Conf. Genetic Algorithms* (San Mateo, California, 1993).
28. J. Horn, N. Nafpliotis and D. E. Goldberg, A niched pareto genetic algorithm for multiobjective optimization, in *Proc. First IEEE Conf. Evolutionary Computation* (Piscataway, NJ, 1994).
29. N. K. Srinivas and A. Deb, Multiobjective optimization using nondominated sorting in genetic algorithms, *Journal of Evolutionary Computation* **2**(3) (1994) 221–48.
30. E. Zitzler and L. Thiele, Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach, *IEEE Trans. Vol. Comput* **3**(4) (1999) 257–71.
31. K. Deb, S. Agarwal, A. Pratap and T. Meyarivan, A fast elitist nondominated sorting genetic algorithm for multi-objective optimization: NSGA-II, *KanGAL Report Number 200001.* (Kanpur, India:Indian Institute of Technology, 2000).
32. K. Deb, S. Agarwal, A. Pratap and T. Meyarivan, A fast elitist nondominated sorting genetic algorithm for multi-objective optimization: NSGA-II, in *Proc. Parallel Problem Solving from Nature VI Conf.* (Paris, France, 2000).

33. K. Deb, A. Pratap, S. Agarwal and T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, *IEEE Transactions on Evolutionary Computation* **6**(2) (2002) 182–197.
34. H. A. Taboada, J. F. Espiritu and D. W. Coit, MOMS-GA: A multiobjective multi-state genetic algorithm for system reliability optimization design problems, *IEEE Transactions on Reliability* **57** (2008) 182–191.
35. L. Fausett, *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications* (Prentice Hall, NJ, 1994).
36. J. Tarja, K. Pekka and W. Jyrki, Structural comparison of data envelopment analysis and multiple objective linear programming, *Management Science* **44**(7) (1998) 962–970.
37. B. H. Demuth, M. Beale and M. T. Hagan, *Neural Network Design* (PWS Publishing Co., Boston, MA, 1997).
38. W. W. Cooper, L. M. Seiford and K. Tone, *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References, and DEA-Solver Software* (Springer, 2006).
39. A. Charnes, W. W. Cooper and E. Rhodes, Measuring the relative efficiency of decision making units, *European Journal of Operational Research* **2** (1978) 429–444.
40. H. A. Taboada, F. Baheranwala and D. W. Coit, Practical solutions for multi-objective optimization: an approach to system reliability design problems, *Reliability Engineering and System Safety* **92** (2007) 314–322.
41. J. W. Yougbare and J. Teghem, Relationship between Pareto optimality in multi-objective 0–1 linear programming and DEA efficiency, *European Journal of Operational Research* **183** (2007) 608–617.

About the Authors

Haitao Liao is an Assistant Professor of Nuclear Engineering Department and Industrial & Information Engineering Department at University of Tennessee at Knoxville, TN. He received his Ph.D. in Industrial and Systems Engineering from Rutgers University in 2004. He also received an M.S. degree in Statistics from Rutgers University. His recent research interests are modeling of accelerated testing, optimal design of testing plans, and maintenance models and spare part logistics. His current research is sponsored by the National Science Foundation and Hong Kong Research Grant Council.

Zhaojun Li is currently a Ph.D. student of Industrial Engineering Department at University of Washington, Seattle, WA. He obtained his M.S. degree in Business Management from Tianjin University in China. He was a Ph.D. student in Industrial and Manufacturing Engineering Department at Wichita State University from 2005 to 2008. His current research interests are reliability engineering and data mining.

A SWNT-Based Sensor for Detecting Human Blood Alcohol Concentration

H. Leng and Y. Lin*

Department of Mechanical and Industrial Engineering , Northeastern University
Boston, MA 02115 USA
yilin@coe.neu.edu

Abstract. Alcohol intake may impair human abilities, degrade human performance, and result in serious diseases. Alcohol sensors are needed to manage the risk and effect of alcohol use to human health and performance. This paper was focused on the theoretical models and design of carbon nanotube based alcohol sensors. The experiments verified that single-walled carbon nanotubes can be used to detect alcohol vapor, and need metal pads to achieve higher sensitivity.

Keywords: Sensor, blood alcohol concentration, carbon nanotube, human-machine system, driver-vehicle system.

1 Introduction and Motivations

Alcoholic beverages are popular in modern society. However, alcohol intake impairs human abilities and degrades human performance [1]. Excessive consumption of alcoholic beverages may result in serious diseases [2]. In order to manage the risk and effect of alcohol use to human health and performance, it is worthy to monitor human blood alcohol concentration (BAC). A widely acceptable method is measuring the alcohol concentration of human exhalation.

Most technologies of measuring alcohol concentration can be classified into three methods, *(1) metal oxide based methods* in which the sensing element is metal oxides such as SnO_2 [3], *(2) optical methods* in which the absorption bands of alcohol are used [4], and *(3) carbon nanotubes (CNT) based methods* in which the resistance of CNTs changes with the ambient alcohol concentration [5]. Compared to the other two methods based alcohol sensors, CNT based alcohol sensors have the potential to achieve ultra-high sensitivity, quick response, large measurement range, compact size, and low energy consumption. These features are essential to monitor human BAC for human health and performance.

This paper is focused on developing a CNT based alcohol sensor which can be used to monitor the alcohol concentration of human exhalation, and then detect human blood alcohol concentration (BAC).

M. Cheng (Ed.): NanoNet 2008, LNICST 3, pp. 69–73, 2009.

A Review of Statistical Methods for Quality Improvement and Control in Nanotechnology

JYE-CHYI LU

Georgia Institute of Technology, Atlanta, GA 30332-0205, USA

SHUEN-LIN JENG

National Cheng Kung University, Tainan 701, Taiwan

KAIBO WANG

Tsinghua University, Beijing 100084, P. R. China

Nanotechnology has received a considerable amount of attention from various fields and has become a multidisciplinary subject, where several research ventures have taken place in recent years. This field is expected to affect every sector of our economy and daily life in the near future. Besides advances in physics, chemistry, biology, and other science-based technologies, the use of statistical methods has also helped the rapid development of nanotechnology in terms of data collection, treatment-effect estimation, hypothesis testing, and quality control. This paper reviews some instances where statistical methods have been used in nanoscale applications. Topics include experimental design, uncertainty modeling, process optimization and monitoring, and areas for future research efforts.

Key Words: Automatic Control; Experimental Design; Nanomanufacturing; Physical-Statistical Modeling; Statistical Quality Control; Stochastic Modeling.

"NANOTECHNOLOGY is the understanding and control of matter at dimensions of roughly 1 to 100 nanometers (nms, [a nanometer equals 10^{-9} meters]), where unique phenomena enable novel applications. Encompassing nanoscale science, engineering, and technology, nanotechnology involves imaging, measuring, modeling, and manipulating matter at this length scale. At the nanoscale, the physical, chemical, and biological properties of materials differ in fundamental and valuable ways from the properties of individual atoms and molecules or bulk matter. Nanotechnology R&D is directed toward understanding and creating improved materials, devices, and systems that exploit these new properties." (National Nanotechnology Initiative (2008)).

In the near future, it is expected that nanotechnology will impact every sector of our economy and daily life. Roco (2004) characterized the development of nanotechnology into four generations, each with different featured products. The first generation, starting from about 2001, is characterized by *passive nanostructures*; the major focus of this generation is on nanostructured materials and tools for measurement and control of nanoscale processes. Popular examples include nanoparticle synthesis and processing, nanocoating, various catalysis, etc. The second generation is characterized by *active nanostructures*, according to Roco (2004). The research focus moves from nanostructured materials to novel devices and

Dr. Lu is a Professor in the School of Industrial and Systems Engineering. His email address is jclu@isye.gatech.edu.

Dr. Jeng is an Associate Professor in the Department of Statistics. His email address is sljeng@mail.ncku.edu.tw.

Dr. Wang is an Assistant Professor in the Department of Industrial Engineering. His email address is kbwang@tsinghua.edu.cn.

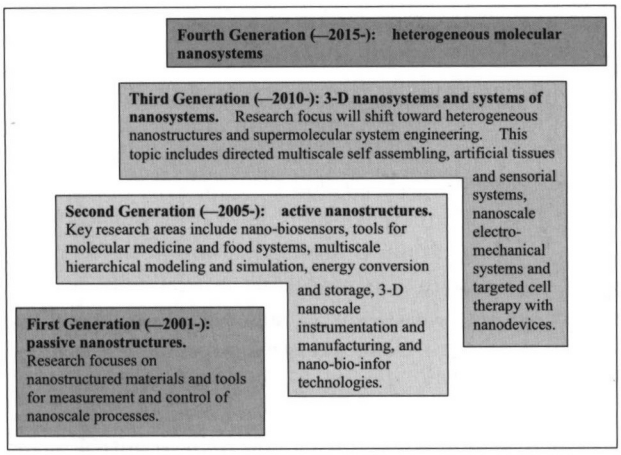


FIGURE 1. Four Generations of Nanotechnology Applications.

device system architectures. Popular research topics include nanobiosensors and nanodevices, nanoscale tools, nanoscale instrumentation, and nanomanufacturing. Modeling and simulation of nanoprocesses are also important topics in this generation. The third generation, featuring *3-D nanosystems and systems of nanosystems*, is expected to shift toward heterogeneous nanostructures and supermolecular system engineering. As an example, research on nanoscale electromechanical systems will be greatly promoted in this stage. Roco (2004) selected *heterogeneous molecular nanosystems* as the feature products of the fourth generation.

Figure 1 presents the overlapping generations of nanotechnology products and their respective manufacturing methods and research focus identified by the International Risk Governance Council (IRGC) (see Roco (2004) for details). We will use this summary to lead the discussion of where statistical methods can contribute (or have contributed) in nanotechnology development in the remainder of this paper.

Various government agencies, private corporations, and venture capitalists have created programs to support R&D in nanotechnology. According to Lux Research, the R&D investments in nanotechnology worldwide amounted to $12.4 billion in 2006. The United States Department of Energy started a nanotechnology program to develop new materials for improving fuel efficiency and providing efficient lighting. Likewise, the Department of Defense has laid out a research plan to develop new approaches and processes for manufacturing novel, reliable, lower cost, higher performance, and more flexible electronic, magnetic, optical, and mechanical devices. Even the research topics from the National Institutes of Health include new nanomaterials for interfacing with living tissues and molecular and cellular sensing for gathering diagnostic data inside human bodies where traditional instruments cannot reach.

Many challenges confronting nanotechnology research call for solutions from a multidisciplinary approach. Because statistical techniques have made sizable impacts in many technology fields in the past, statistics are expected to play an important role in tackling these challenges and boosting the development of nanotechnology. The purpose of this paper is to present examples of applications of statistical methods in nanotechnology research and to identify possible new statistical approaches for solving emerging problems in nanotechnology. Specifically, we highlight the following challenges that provide opportunities for statistical applications:

First, *statistical procedures help us learn more about the formation processes of nanocomposites.* Recently, researchers have reported new progress in developing nanocomposites. Forming processes dictate properties of nanocomposites. These processes usually involve complex chemical and mechanical reactions, where even minor changes of environmental factors or process settings may result in unexpected outcomes. Because the theory of the nanocomposite-forming process has not matured yet, many developments rely on experimental studies. Due to the high costs associated with experimental runs and sample measurements and the complex relationship between process outcomes and controllable/noise factors, new statistical experimental design methods are needed. Section 2 reviews recent work on applying experimental designs in nanocomposites applications and categorizes them based on the types of designs adopted in each study.

Second, *statistical modeling helps deal with special data types and processes.* Data collected from processes for developing nanoscale materials sometimes exhibit forms that are different from the usual patterns seen in conventional manufacturing situations. Such characteristics may include experimental spaces with many isolated regions of low or zero yields and high-frequency signals shown in spatial domains. Therefore, extensions of commonly used statistical techniques are required to draw useful information from such types of data. Beyond these extensions, stochastic modeling of the forming processes for characterizing multilevel or multiscale uncertain-

ties is another valuable topic. Section 3 summarizes recent work on data collection, statistical analysis, and uncertainty modeling in nano research and provides examples for researchers that are facing similar problems.

Third, *statistics help improve low-quality and high-defect processes.* Because nanosyntheses and nanofabrications are not well controlled thus far, defect rates of nanocomposites remain rather high. Mass and high-yield production is the key step confronting nanomanufacturing research. Statistical quality control and productivity improvement techniques, which proved to be tremendously helpful in traditional manufacturing, are needed in nanomanufacturing to enhance product quality and improve production efficiency. Extension of the traditional quality monitoring and improvement tools to accommodate the potential flood of sensor information is an interesting statistics research topic. Section 4 reviews the use of statistical process control (SPC) and automatic process control (APC) techniques in nanoapplications. The feedback control section provides examples of using stochastic partial differential equations (SPDEs) to describe thin-film deposition processes. Based on these SPDEs and results in kinetic Monte-Carlo simulations, multiscale automatic process controllers (APCs) are developed.

Last, *statistics help improve low and unpredictable product reliability.* Reliability models for nanocomposites are not well developed. However, through investigation of nanomaterials interfacing with environmental and stress factors, the reliability characteristics of nanodevices can be better understood and predicted. Jeng et al. (2007) provide a comprehensive review on recent reliability studies in nanoapplications. To keep this article brief, reliability issues will not be discussed in this review.

The remainder of this paper is organized as follows. Section 2 reviews successful applications of design of experiments (DOEs) techniques in nanotechnology research, which covers regular, nonregular, and robust parameter designs. Section 3 reviews various data collection, statistical analysis, and modeling methods in the literature. Specifically, works on probability distribution and variation modeling of nanostructure characteristics, treatment of high-frequency signal and spatial data as well as stochastic modeling techniques are presented. Section 4 reviews recent development of SPC and quality-control techniques. Examples of process monitoring and feedback control, especially multiscale modeling and control techniques, are illustrated in this section. Section 5 concludes this review with suggestions for future research on quality improvement and control in nanotechnology.

Design of Experiments

Due to the applications in nanoelectronics, photonics, data storage, and sensing, synthesizing nanostructures is a research topic of foremost importance in nanotechnology (Dasgupta et al. (2008)). However, the synthesis process is extremely sensitive to control settings and environmental noises. For example, to generate the nanostructures of nanosaws, nanowires, and nanobelts shown in Figure 2, Dasgupta et al. (2008) shows that the control factors, such as temperature and pressure, have a heavy impact on the final output of a synthesis process. Moreover, situations occur in which multiple responses are studied with functional relationships containing

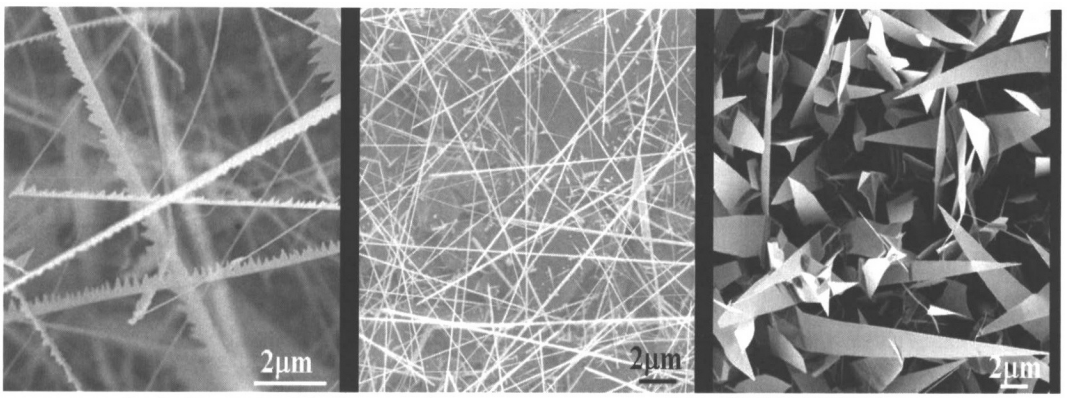


FIGURE 2. Nanosaws, Nanowires, and Nanobelts (from Dasgupta et al. (2008)).

many potential factors of interest. In order to robustly optimize several properties of nanoproducts, the impact of each factor, plus their interactions, on process outputs should be studied using advanced statistical techniques. In the literature, DOE has been employed as the major tool for exploring the relationship between controllable/noise factors and process responses. Regular designs, including full factorial and fractional factorial designs, response surface methodologies, and nonregular designs (e.g., D-optimal designs) have been successfully utilized to optimize synthesis processes and investigate material properties. This section reviews the use of DOE techniques and robust parameter designs in nanotechnology and gives examples of how these techniques can be utilized in the nanomanufacturing processes. This review covers studies in the following journals: *AAPS PharmSciTech*; *Carbon*; *Chemical Engineering Journal, Colloids and Surfaces A: Physicochemical Engineering Aspects*; *International Journal of Pharmaceutics*; *Journal of Physical Chemistry B*; *Journal of the American Statistical Association*; *Materials and Design*; *Microelectronics Reliability*; *Nanotechnology*; and *Powder Technology*.

Regular Designs

Basumallick et al. (2003) investigated the synthesis processes for Ni–SiO_2 and Co–SiO_2 nanocomposites, of which the physical properties are very sensitive to process parameters. This attribute makes the response surface very rugged. Because processing conditions have significant impact on the properties of nanocomposites, the authors considered three factors with three levels each in their experimentation. Instead of using a conventional three-level fractional factorial design, the authors designed eight runs using a two-level full factorial design and added three runs setting all factors at the middle level. Using regression equations fitted by the 11-run experiment outcomes, the fractional conversion values of the nanocomposites were modeled as a function of heating rate, composite concentration, and the starting temperature of the reaction. Becaise the response surface is very rugged, in order to improve the data collection and analysis methods used in this paper, we recommend the experimental designs (and their extensions) commonly used in computer experiments (e.g., Dasgupta et al. (2008)) for efficiently collecting costly data. Moreover, robust process-optimization ideas (e.g., Taguchi (1986)) could be used so that the physical properties of the nanocomposites are less sensitive to the noise factors.

Lin et al. (2003) studied the surface and grain structure of silver-plated film on a copper lead frame. The surface roughness was measured by atomic force microscopy (AFM), while the surface thickness was obtained by nanoindentation measurement using a UMIS-2000 nanoindenter. Additionally, the grain structure was measured by a transmission electron microscope (TEM). The impact of the silver-plated film-surface characteristics and grain structures to the quality of wedge bonding between gold wire and silver-plated lead frame were investigated through a 2^4 factorial design. Four design parameters were bond time, bond force, electrical current, and the two lead frame types. The experimental results showed that the surface characteristics and grain structures have a great impact on the quality of bonding.

To systematically obtain a model of factors leading to an optimum response, many researchers use the response surface methodology (RSM) for data collection and process optimization. RSM should also be considered when complex models are needed for characterizing a nanosynthesis or nanomanufacturing process. Prakobvaitayakit and Nimmannit (2003) presented a process to prepare polylactic-co-glycolic acid (PLGA) nanoparticles by interfacial deposition following the solvent-displacement technique. They investigated the process through a 2^3 factorial design experiment. Multiple responses—particle size, amount of encapsulated material, and encapsulation efficiency—were considered in the experiments. To optimize multiple responses, the authors used a simultaneous optimization technique with a desirability function. The desirability function converts each response into an individual function that varies over the range [0, 1] and takes the value one when the response is at its target value and less than one if not (see page 425 in Montgomery (2005) for details). By defining the overall desirability as the product of all individual functions, optimizing multiple responses is achieved by maximizing the overall desirability. With this method, the experimental design helped choose the optimal formulation ingredients for the nanoparticles. Now, the composites formed by PLGA nanoparticles are being commercially used for drug-delivery systems.

Barglik-Chory et al. (2004) used a second-order response surface to study the main effects and interactions of three controllable factors in synthesizing a type of semiconductor nanoparticles, colloidal CdS. The quantum effects in nanoparticles provided discrete energy levels, and the semiconductor band gap exhibited strong size dependence. Through con-

trolled experiments, the influence of environmental factors on nanoparticles could be investigated and quantified, potentially leading to an improvement in the nanofabrication efficiency.

After identifying the significant parameters by using Taguchi's parameter design method (Taguchi (1986)), Hou et al. (2007) used RSM to build a relationship between five process parameters and average grain size of nanoparticles. They found that the relationship between process parameters and grain size can be modeled with a second-order equation. Using the integrated genetic algorithm and RSM approach, the optimal settings of these five parameters in the nanoparticle milling process were determined.

Yördem et al. (2008) used RSM to investigate effects of material choices and process parameters on the diameter of electrospun polyacrylonitrile nanofibers. Three explanatory factors—voltage, solution concentration, and collector distance—and two response variables—fiber diameter and coefficient of variation—were studied. At each level of the collector distance, the other two factors were varied and the resultant fiber diameter was recorded. For the purpose of predicting fiber diameter and coefficient of variation, polynomial regression models were fitted to the experimental data at each level of the collector distance. Interactions among the factors were also identified. The authors suggested a narrower window of factor space for future nanofiber production.

Nazzal et al. (2002) used a three-level Box–Behnken design to evaluate the effect of formulation ingredients on the release rate of ubiquinone from its adsorbing solid compact and to obtain an optimized self-nanoemulsified tablet formulation. To study the effects of three independent variables on six responses, 15 runs of experiments were conducted. The formulation ingredients—copolyvidone, maltodextrin, and microcrystalline cellulose—were shown to have a significant effect on the emulsion release rate. However, this article did not show how to allocate these ingredients to optimize the multiple objectives.

Using a radio-frequency plasma reactor, Cota-Sanchez et al. (2005) used a 2^4 factorial design to study fullerenes synthesis. The response variable was C_{60} yield and the four operating parameters were reactor pressure, raw-material feed rate, carbon black–catalyst ratio, and generator-plate power. Using ultraviolet spectrometric analysis to measure yield, they found that the significant factors that affect the C_{60} synthesis are the reactor pressure, plate power, and feed rate. Under the optimal operating condition, the yields were synthesized up to about 7.7 wt.%; moreover, nanotubes were successfully produced. Further optimization of other parameters, such as the particle injection and quenching conditions, through the use of RSM might lead to enhanced yields of these carbon nanostructures.

In order to study the flexural properties and dispersions of a ceramic material, SiC, Yong and Hahn (2005) employed a two-level full-factorial design to investigate interactions between coupling agent and dispersant. A central composite design was used to determine the optimal dosages of chosen factors to achieve the maximum flexural strength and maximum particle dispersion. Experimental results showed that two objectives could be optimized simultaneously at the same factor levels. When an optimal dosage is employed, the nanoparticle reinforcement can enhance the mechanical properties of the composites.

Compared with factorial designs, split-plot designs are suitable for situations in which physical restrictions on a process exist and certain combinations of factor levels are difficult to reach. Nembhard et al. (2006) presented a split-plot design for investigating nanoscale milling of submicron channels on a gold layer. Similar to a synthesis process that involves complex chemical and mechanical reactions, such a milling process was sensitive to various controllable and noise factors. By treating the experiment as a two-stage process, whole-plot and subplot factors were identified. Results showed that split-plot experiments in nanomanufacturing reduce labor and costs and were often effective at detecting the effects of subplot factors.

In order to achieve high yield and reproducibility of a synthesis process, Dasgupta et al. (2008) conducted a full-factorial experiment. Two parameters, temperature and pressure, with five and nine levels, respectively, were varied during the experiment and three response variables, the numbers of nanosaws, nanowires, and nanobelts, were recorded. Because the total number of the nanostructures was a constant, the authors proposed simultaneously modeling the probability of generating different nanostructures using a multinomial generalized linear model (GLM). The optimal settings obtained from this research led to a significant improvement in the process yield.

Nonregular Designs

When a regular factorial design is not feasible due to possible limitations in experimental run-size and factor-level selections, some nonregular designs may be considered for nano process modeling and optimization. Among others, successful applications of D-optimal designs in nanotechnology have been demonstrated by several researchers. Compared with the regular designs, D-optimal designs may use nonorthogonal design matrices to reduce experimental runs and minimize the variances of coefficients associated with a specific model setting.

Fasulo et al. (2004) studied the extrusion processing of thermoplastic olefin (TPO) nanocomposites. Properties of TPO composites are influenced by the forming process. The authors chose different combinations of processing factors, including melt temperature, feed rate, and extruder screw-rotation speed, and investigated both surface appearance and physical/mechanical properties of the nanocomposites. A D-optimal design was adopted in the research to characterize the relationship between the quality measures and explanatory factors. Useful suggestions for optimizing process output were obtained.

Dasgupta (2007) developed a sequential minimum energy design (SMED) to deal with complex response surface with multiple optima for the yield of nanomaterial synthesis by using fewer design points. Figure 3 illustrates the response surface of the yield and the selected design points. Compared with traditional designs, the SMED design can probe high-yield regions and avoid nonyield points more effectively.

Robust Parameter Design

One major challenge confronting nanosynthesis/manufacturing processes is the high variation in experimental results. Most synthesis processes are very sensitive to environmental or noise factors. Therefore, robust parameter design (e.g., Taguchi (1986)) has been considered by several researchers to reduce experimental variation and enhance process yield and production efficiency. Figure 4 illustrates one of the key ideas of the robust parameter design. The level setting of control factors will be different by including the noise factors in the experiment when there are interactions between control and noise factors.

Nanoparticles have been widely utilized in many industrial applications, such as carbon nanotube, nanoceramics, and nanocompound materials. The wet-type milling machine is a recently developed tool to produce nanoparticles and to avoid aggregation effect. Because of its simplicity and applicability to all classes of materials, this machine is becoming popular. Hou et al. (2007) applied the robust parameter design method to optimize a nanoparticle milling process. They considered the following five process factors, each with three levels, to improve the nanoparticle milling process: the milling time, flow velocity of circulation systems, rotation velocity of agitator shaft, solute-to-solvent weight ratio, and filling ratio of grinding media. The response variable was the nanoparticle grain size, which is measured by the Coulter multisizer equipment. To save experimental cost, the L_{27} orthogonal array was used. The experimental results showed that all five process factors significantly affect the grain size.

Kim et al. (2004) implemented the robust design method with an L9 orthogonal array to optimize the recipe for preparing nanosize silver particles. The silver nanoparticles have been widely used in chemical and medical industry due to their unique properties of conductivity and resistance to oxidation. The objective was to determine the experimental conditions where the size of nanoparticle is small and has less variability. The following three factors were considered in the experiment: molar concentration ratio, dispersant concentration, and feed rate. The response variables are the average size and the size distribution of silver nanoparticles. The concentration of dispersant was identified as the most influencing factor on the average particle size and the size distribution. Using the derived optimal conditions, silver nanoparticles can now be prepared with small size variance using the derived optimal condition.

In another paper about the synthesis of nanoparticles, Kim et al. (2005) used Taguchi's method to optimize a new microemulsion method for preparing TiO_2 nanoparticles. An L_8 orthogonal array was used as the design of experiment for the five factors: H_2O surfactant value, H_2O/TEOT value, ammonia concentration, feed rate, and reaction temperature. The derived optimal condition led to the least size variability in nanoparticles.

Data Collection, Statistical Analysis, and Physical–Chemical–Statistical Modeling

Because measurements with complicated data patterns are frequently seen in experiments in nanotechnology, the collection, analysis, and modeling

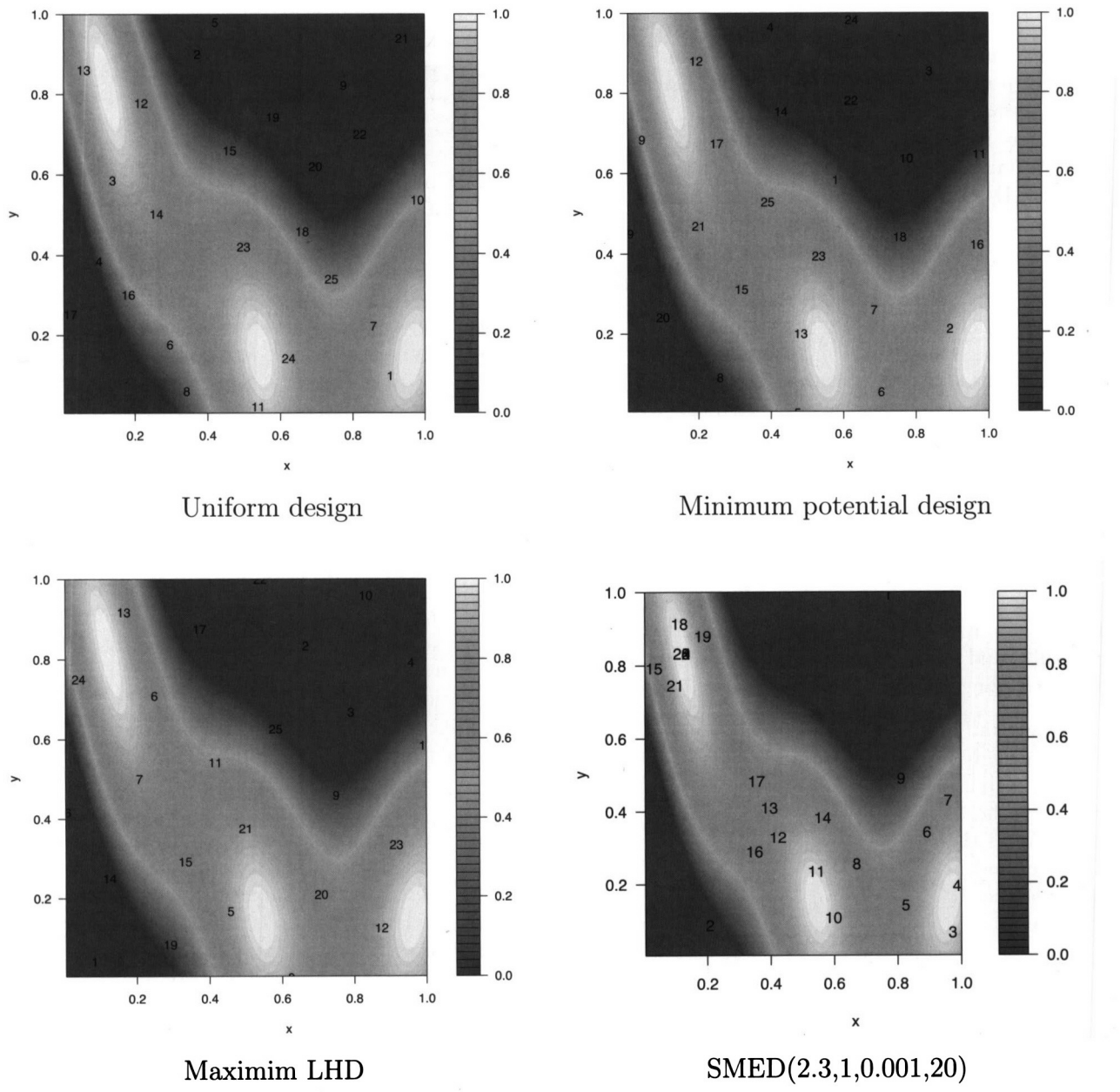


FIGURE 3. Response Surface of the Yield of a Nanostructure and the Selected Design Points (from Dasgupta (2007)).

of data in nanoengineering are vital topics. This section focuses on high-frequency and/or spatial data signals seen in nanotechnology studies. In addition, probability and stochastic models of nanoscale measurements and process variations are reviewed. The primary journals reviewed here include: *Combustion Theory and Modeling, IEEE Transactions on Nanotechnology, IEEE Transactions on Reliability, IEEE Transactions on Semiconductor Manufacturing, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, IEICE Transactions on Electronics, International Journal of Engineering Science, International Journal of Plasticity, Journal of the Mechanics and Physics of Solids, Physical Review Letters, Solid-State Electronics*, and *Surface & Coatings Technology.*

Sampling Plans

Effective sampling plans are critical to model the nanofabrication processes using fewer data. The fol-

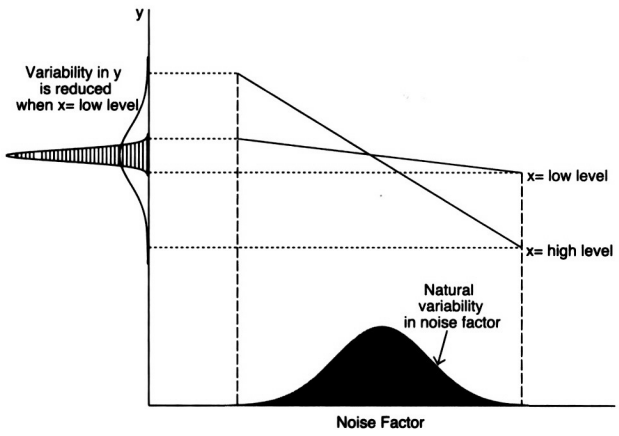


FIGURE 4. One of the Key Ideas of the Robust Parameter Design.

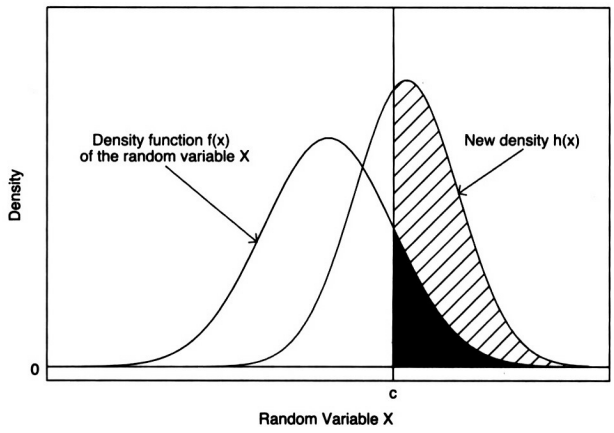


FIGURE 5. Idea Illustration of Importance Sampling.

lowing sampling techniques can be used to select representative data for analysis: simple random sampling, stratified sampling, cluster sampling, systematic sampling, importance sampling, and their hybrids. However, new sampling techniques have been developed for the nanofabrication process.

Zhao et al. (2004) developed a new double sampling technique to test interconnects and buses in very large scale integration (VLSI) circuits. Due to the rapid size scaling and high-speed operation of integrated circuits (ICs) using nanometer technologies, electromagnetic noise sources and their effects on interconnects have become extremely significant. The basic idea for this sampling technique was to sample test data for loading into two flip-flops at a fixed time interval and check whether the resultant data streams were consistent with each other. Thus, inconsistent and noisy signals with spikes could be captured. This technique was capable of detecting errors caused by electromagnetic noise effects in nanofabrication.

Chang et al. (2005) proposed a new method to detect the read failure in a static random access memory (SRAM) cell using a critical-point sampling technique. The idea was similar to the importance sampling. In sub–100-nm technologies, the analytical model previously used fails to accurately match realistic simulation results due to various short channel effects and different leakage components. It is preferable to employ a full-scale transient Monte Carlo (MC) simulation; however, using the MC method usually takes a large number of iterative runs to obtain an accurate read failure probability. Thus, rather than deriving the entire voltage-transfer characteristic curve, the authors measured the SRAM cell stability at certain representative points on the curve for a specified voltage value. The experimental result showed that their model achieves high accuracy and was 20 times faster in computational speed.

Proper use of importance sampling will provide a more accurate inference or save the sampling cost with smaller sample size. This is a very helpful statistical technique for collecting expensive samples in many nanofabrication processes. Figure 5 illustrates the general idea. Suppose the target of the inference is the mean of a statistic $m(X)$, i.e. $E(m(X))$, where $m(x)$ only depends on the sample values x greater than a constant c. Then a proper choice of a new density $h(x)$ of X will make the sample mean of $m(X)f(X)/h(X)$ an unbiased estimator of the target with smaller variance. The main character of the new density $h(x)$ is to produce more samples on the area (greater than c) that affects the values of $m(x)$. See page 87 in Davison (2003) for more details of importance sampling.

Probability Distribution and Variation Modeling

As technology shrinks to the nanoscale region, probability distributions for material and structure performance may change to different forms. Moreover, variability in device performance becomes a major issue in the circuit-design stage. The following paragraphs review several papers studying probability distributions of nanoparticle measurements and process variations in IC manufacturing.

In ordnance technologies, the reproducibility of a

thermite's burning behavior is critical for both safety and performance evaluation. As a particle's diameter approaches the nanoscale, the burn-rate calculation becomes increasingly sensitive to variations in the particle diameter. In the study by Granier and Pantoya (2004), the burn-rate estimates for nanoscale thermites were statistically evaluated with a probability density function (pdf) of the particle-size distribution and a diameter-dependent burn-rate equation. Based on a series of scanning electron microscopy (SEM) images, a model of mass fractal aggregates was used to interpret the scattering data. A volume-weighted particle-size distribution was obtained. Both single mode and bimodal particle-size distributions were studied using Gaussian and a mixture of Gaussian distributions, respectively. The analysis showed that, as the particle size reduced to the nanoscale, the size distribution, rather than the average particle size alone, became increasingly important. Large variability in the burn rate was associated with a large standard deviation in particle sizes.

Bazant and Pang (2007) studied the pdf of strength of nanostructures based on a nanoscale atomic lattice. To predict a failure event with an extremely low probability, the cumulative distribution function (cdf) of strength of quasi-brittle structures was modeled as a chain of representative volume elements (RVE). Each of the RVEs was statistically characterized by a hierarchical model consisting of bundles (via a parallel coupling) of two long subchains. Each of the subchains consisted of subbundles of two or three long sub-subchains of sub-subbundles and so on. Eventually, the nanoscale atomic lattice was reached. They gave physical reasons that the failure of interatomic bonds follows a thermally activated process governed by stress-dependent activation energy barriers. The tail of the cdf of the RVE strength followed a power-law model. Finally, they concluded that the distribution of strength of the quasi-brittle nanostructure is a Weibull.

The physical process of surface formation in nanocomposites has several stochastic components. Chen et al. (2005) studied barrier effects impacting surface formation using oxidized porous silicon. According to their experimental results, a Gaussian distribution fit the data of oxidized porous silicon samples well. The barrier height turned out to have a Gaussion distribution, and the photon energy was shown to be a function of the barrier heights.

Hydrogenated amorphous silicon (a-Si:H), the prototypical disordered semiconductor, is an important material for use in nanoscale optoelectronic applications including sensors, flat-panel displays, 2D medical imaging, and photovoltaics. Belich et al. (2003) reported that the noise distribution in a-Si:H was non-Gaussian in nature, a surprising feature in macroscopic samples at and above room temperature. Traditionally, the noise arose from an ensemble of statistically independent fluctuators. This led to Gaussian-distributed noises. One possible explanation of the non-Gaussian noises was due to (nonlinear) interactions between fluctuators.

The stored charge in each logical node of a VLSI circuit decreases with decreased dimensions and decreased voltages in nanotechnology. Weak radiation can cause disturbance in the circuit signals, leading to increased soft-error failures. The fault/error-detection probability can serve as a measure of soft-error failure, which is an increased threat in the nanodomain logic node. In order to model the fault/error-detection probability, Rejimon and Bahnja (2005) presented a Bayesian network for error-sensitivity analysis in VLSI circuits. Their procedure provided a framework to obtain the joint pdfs of all variables in the network for calculating the error-detection probability.

As the technology node goes down to 90 nm and below, variability in device performance becomes a crucial problem in the design of ICs. In the past, die-to-die variability, which was well managed by the worst-case design technique, dominated the within-die variability. Onodera (2006) pointed out that the statistical nature of the variability has been changed such that the within-die variability is growing. This characteristic presents a challenge in circuit-design methodology. His paper provided measurable results of variability in 0.35-, 0.18-, and 0.13-μm processes and explained the trend of variability. It also showed that a circuit that was designed optimally under the assumption of deterministic delay is now most susceptible to random fluctuation. This result indicates the need of applying statistical thinking in the circuit-design methodology.

For solving process-variation problems in nanoscale-IC manufacturing, Li et al. (2005) proposed a novel projection-based extraction approach, PROBE, to efficiently create quadratic response surface models and capture both interdie and intradie variations with affordable computation cost. Instead of fitting a full-rank quadratic model, PROBE applied a projection operator and found an optimal

low-rank model by minimizing approximation errors, which were defined by the Frobenius norm. In PROBE, the modeling accuracy and parsimony can be tuned by increasing or decreasing the dimension of the projection space. Several examples from digital and analog circuit-modeling applications demonstrated that PROBE can generate accurate response surface models while achieving up to 12 times faster speed as compared with traditional process-variation simulation and modeling methods.

Stochastic Modeling

Due to the fact that a large proportion of variability in nano synthesis/growth processes are not well explained by known physical models, stochastic modeling techniques have served as an effective way in characterizing such processes.

Miranda and Jimenez (2004) proposed a stochastic logistic model based on a Wiener process for characterizing the breakdown dynamics of ultrathin gate oxides (at 2-nm level) and for understanding the effect of voltage stress on the leakage current. The model had two components: a deterministic term with a logistic function describing the mean failure behavior and a random term with a Wiener process representing uncertainties and noises. Throughout the model, the nano-material–degradation dynamics were captured using a small set of parameters.

To characterize degradation of leakage currents of 3-nm gate oxides, Hsieh and Jeng (2007) considered a nonhomogeneous Weibull compound Poisson model with accelerated stresses. The oxides were in square metal-oxide semiconductor (MOS) capacitors grown on p-Si. One hundred twenty capacitors were irradiated at three levels of ion density, and the leakage current versus gate voltage was measured before and after each irradiation step. Figure 6 shows the sketch of the leakage current of the gate oxides at three ion levels and two accelerated voltages. They provided maximum-likelihood estimates of model parameters and derived the breakdown-time distribu-

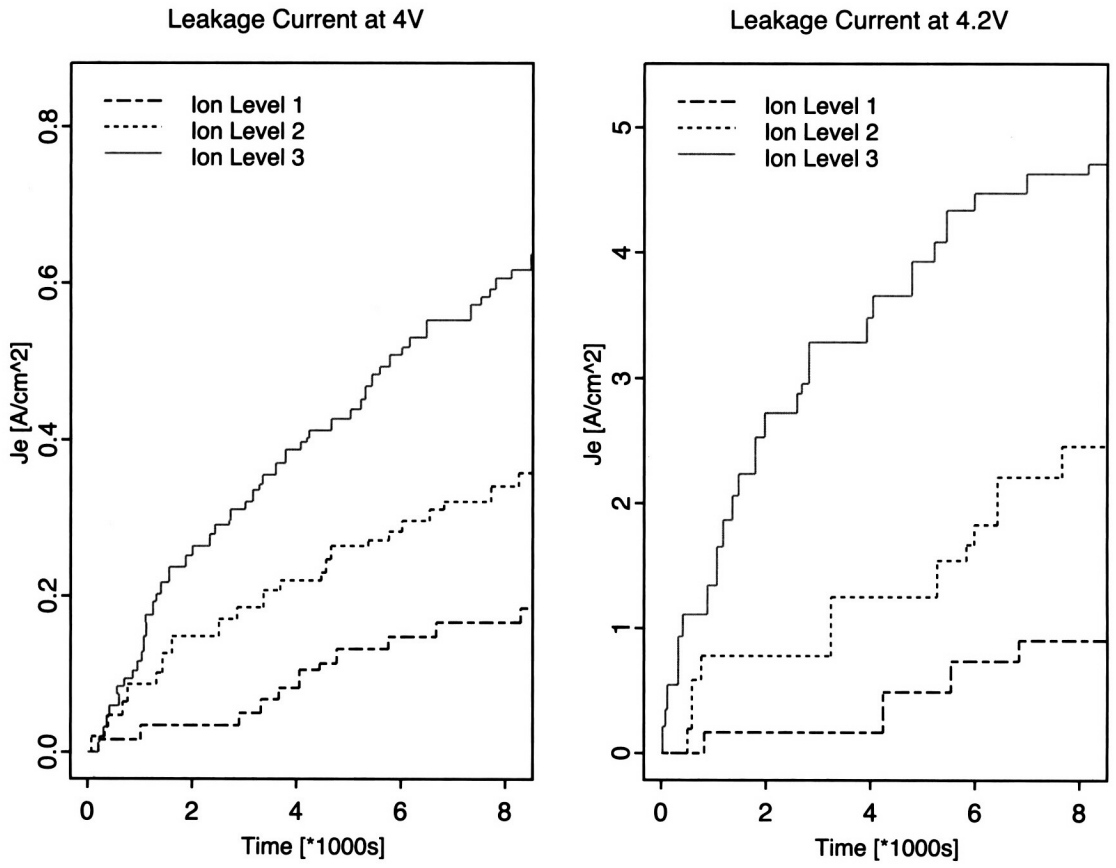


FIGURE 6. Leakage Currents of 3-nm Gate Oxides Under Accelerated Stresses (Voltages; from Hsieh and Jeng (2007)).

tion. To check the proposed models for the degradation measurements and the rate of breakdown-event occurrence, goodness-of-fit tests were considered. The estimated nanodevice reliabilities were calculated at lower stress conditions.

Time series of nanosystem measurements exhibit intermittency. At random times, the system switches from state *on* (or up) to state *off* (or down) and vice versa. Margolin and Barkai (2005) investigated the nonergodic properties of blinking nanocrystals modeled by a Levy-walk stochastic process. The process was characterized based on the sequence of on and off sojourn times. The times of the on–off events were mutually independent and were drawn at random from a pdf that followed a power-law distribution generated by a fractional Poisson process.

It has been widely recognized that novel nanoelectronic devices, such as carbon nanotubes and molecular switches, have high manufacturing defects due to the stochastic nature of the bottom-up physical and chemical self-assembly processes. Qi et al. (2005) reported that it was very challenging to manufacture nanocomputers with a device density of 10^{12} chips due to faulty components pervasive in the device. The authors studied the behavior of a NAND-multiplexing system with a Markov chain of distributions that could be unimodal or bimodal, depending on whether the probability of NAND gates was larger or smaller than a threshold value.

High-Frequency Signal and Spatial Data Analysis

Due to the use of advanced data-collection equipment, spatial data and/or high-frequency signals are collected in many nano applications. Because such data signals contain rich information about process status or product quality, effective (and efficient) statistical analysis methods are greatly important for extracting knowledge. The following paragraphs present a few examples.

Spatial statistical models have been used in modeling nanoscale structures. Chen and Lee (2004) used lattice points to represent atomic bonding units in a crystal. The structure of the unit together with the network of lattice points determined the crystal structure and the physical properties of the nanomaterial. In their experiments, polycrystalline solids consist of randomly distributed grains and grain boundaries. The size of grains was usually in the nano/microscale. Each grain was modeled as crystallized solid by micromorphic theory, while the grain boundaries were modeled by classical continuum theory. Within each grain, the atomic motion led to the continuum lattice deformation from nanoscale to microscale. A multiscale spatial model was then used to characterize the material behavior of polycrystalline solids.

In order to understand the macroscopic properties of heterogeneous materials, modeling spatial correlations for microstructures is needed. Jefferson et al. (2005) analyzed microstructures of polymer nanocomposites and discovered that the material measurements did not exhibit randomness from a single homogeneous distribution. The traditional empirical model became inappropriate. A mixture of two distributions that defined the transition probabilities between two phases of the composites was introduced as a tool to examine the randomness and periodicity in the microstructure.

For manufacturing high-performance films with a thickness of only a few nanometers, nondestructive testing methods are required to ensure production quality. Schneider et al. (2002) developed a laser-acoustic technique based on surface acoustic waves for testing nanometer film's hard-coating quality. A nitrogen-pulse laser was used to generate a wide-band surface acoustic wave. When propagating through the nanofilm surface, the wave signals were recorded using a digitizing oscilloscope. Applying the Fourier transformation to the high-frequency laser-acoustic signals, the authors obtained empirical estimates of thickness and hardness of the films. Because the recent popular tool for modeling high-frequency signals is wavelets, it might be interesting to see the applicability of wavelets and the results they lead to in analyzing laser-acoustic signals. See papers in the subsection of process monitoring of functional data for examples.

Statistical Process Control (SPC) and Quality Control

As a traditionally popular technique in manufacturing, SPC is expected to serve the role of identifying assignable causes and thus reducing variations in nano-manufacturing processes. However, in this early stage of research in nanotechnology, nano-production systems have not matured yet. Hence, publications on SPC application to nano-technology research are scarce. This section highlights two important issues appearing in nano technology: functional data monitoring and feedback control. Although most of the feedback control studies deal with control-theory,

i.e., they are not statistically oriented, this review does dig out a few sensor-data–based control studies. In addition, several multiscale modeling and control publications illustrate the potential of future statistics research in the important nanotechnology development areas (see the description in the second generation of Figure 1). This review covers the latest progress on nano-SPC published in the following journals: *Computers & Chemical Engineering, Control Engineering Practice, IEEE Transactions on Plasma Science, IEEE Transactions on Semiconductor Manufacturing, Journal of Process Control, Journal of Quality Technology, Nanotechnology, Surface & Coatings Technology,* and *Wear.*

Process Monitoring of Functional Data

Functional data characterize quality or reliability performance of many manufacturing processes. They are very informative in process monitoring and controlling for nanomachining, ultrathin semiconductor fabrication, and several other manufacturing processes seen in the literature. In particular, wavelet analysis has been popular in modeling and monitoring functional data.

Ganesan et al. (2003) combined online detection with offline modeling strategies in the nanodevice wafer-polishing process. Through wavelet-based multiresolution analysis methods, the delamination defects were identified by analyzing the nonstationary sensor signals based on acoustic emission and coefficient of friction. Jiang and Blunt (2004) introduced a wavelet model to extract linear and curvelike features from complicated nonstationary surface-texture patterns in nanometer morphological structures. Through wavelet decompositions and reconstructions, Das et al. (2005) removed noises from the coefficient of friction signals in their chemical mechanical polishing of nanodevices and used a sequential probability ratio test to set up a control chart for monitoring process conditions.

Feedback Control

Sensor-Data–Based Control

Due to the ultrasmall size of objects and very fine manufacturing processes, using data collected from various sensors for developing precise automatic control rules is critical in nanotechnology studies. For instance, Ohshima (2003) pointed out that most traditional feedback controllers using existing sensors for micro and larger systems were not effective for dealing with nanomachines, where physical and chemical phenomena occur in very short time windows. The author considered *in-situ* feedback schemes to control physical–chemical reactions in their material processing courses. Klepper et al. (2005) developed an *in-situ* feedback–control tool to improve the reproducibility of a nanostructure deposition process. In order to reduce variability in the coating process, the investigators needed to find ways to measure process variations. Experiments were set up to test the assumption that plasma discharge was sensitive to the surface-roughness variation. The experiments resulted in the identification of the correlation between hydrogen atomic emission and formation of metal-carbide coatings. Through the control of the emission, variability in the nanostructure composition and the wear performance of the coating was controlled by a closed-loop feedback algorithm.

Lin et al. (2003) studied precise positioning issues in a nanoscale drive system. To guarantee the desired precision, the authors used an integral type of controller in the positioning. The study demonstrated the possibility of achieving a very precise positioning at the resolution level of sensor measurements. Lantz et al. (2005) introduced a novel micromachined silicon-displacement sensor with displacement resolution of less than 1 nm for providing accurate positioning information. This new technique can detect the displacement by measuring the difference between the resistances of the two sensors.

Multiscale Modeling and Control

A few factors, such as the following, motivate the current active research on control of multiscale processes (see the preface of the special issue on control of multiscale and distributed process systems in 2005, volume 29, *Computers and Chemical Engineering* for details).

(a) key technological needs in semiconductor manufacturing, microtechnology, nano technology, and biotechnology; and

(b) recent developments in actuator and sensor technology that make control of material microstructures, spatial profiles, and product-size distributions feasible and practical.

Using the thin-film growth process as an example, Christofides and Armaou (2006) provided a review of recently developed methods for controlling multiscale processes. The typical thin-film growth process has been widely used in microelectronic devices, nanomaterials, and nanocomposites (see, e.g., Ng et al. (2003), Sneh et al. (2002), Mae and Honda (2000)).

To achieve better control of multiscale processes, process modeling and prediction by using physical, statistical, or simulation methods to characterize complicated process outputs becomes a critical issue. The following provide a few examples.

Precise control of film properties in nanoscale semiconductor manufacturing requires models that predict how the film state (in the microscopic scale) is affected by changes in controllable parameters (in the macroscopic scale). Lou and Christofides (2003) developed a multiscale model that involves coupled partial differential equations (PDEs) for modeling the gas phrase of the material deposition process. Subsequently, the authors used a kinetic Monte-Carlo (kMC) simulator to model the atom adsorption, desorption, and surface-migration processes for shaping thin-film microstructures. There are strong interactions between the macro- and microscale phenomena. For example, the concentration of the precursor in the inlet gas governs the rate of adsorption of atoms on the surface, which, in turn, influences the surface roughness. On the other hand, the density of the atoms on the surface affects the rate of desorption of atoms from surface to the gas phrase, which, in turn, influences the gas concentration of the precursor.

Lou and Christofides (2003) constructed an efficient estimator for the surface roughness at the timescale comparable to the real-time evolution of the process using discrete on-line measurements. Then, the estimated roughness was fed into a proportional-integral (PI) controller. Application of the estimator/controller to the multiscale process model demonstrated successful regulation of the surface roughness at the desired value.

In nanoscale semiconductor manufacturing, the kMC simulation methods have been used for

1. predicting microscopic properties of thin films (e.g., surface roughness);
2. studying the dynamics of complex material deposition processes including multiple chemical species with both short-range and long-range interactions; and
3. performing predictive control design to control final surface roughness.

kMC is not available in closed form, thus making it difficult for use in system-level analysis and design and implementation of model-based feedback-control systems.

Instead of using the kMC approach in developing feedback-control schemes, for many deposition and sputtering processes, a system of stochastic linear/nonlinear PDEs is derivable based on microscopic rules corresponding to the so-called master equation, which describes the stochastic nature of the thin-film growth process (Van Kampen (1992)). Lou and Christofides (2006) proposed using nonlinear stochastic ordinary differential equations (ODEs) (an approximation of the stochastic PDEs) for developing computationally efficient feedback controllers. Their objective was to control the expected roughness of the surface. The solution strategy is to control the covariance of the thin-film growth states in the nonlinear stochastic PDE system for various spatial locations. Based on various simulation runs, it is clear that the proposed nonlinear feedback-control method can reduce the surface roughness to the desired level, while also effectively reducing the variance of the final surface roughness.

Different control strategies have been used to control multiscale processes. Both proportional-integral (PI) controllers (Lou and Christofides (2003), Lou and Christofides (2006), Gallivan (2005)) and model-based predictive controllers (Ni and Christofides (2005a, b), Christofides and Armaou (2006)) have been implemented in practice to control thin-film growth surface roughness and other parameters of interest. Figure 7 shows a general framework of predictive controllers. The difference between the real and predicted output are compared; such information is then fed into the optimizer to generate an optimal set point for the next step. Multiscale models

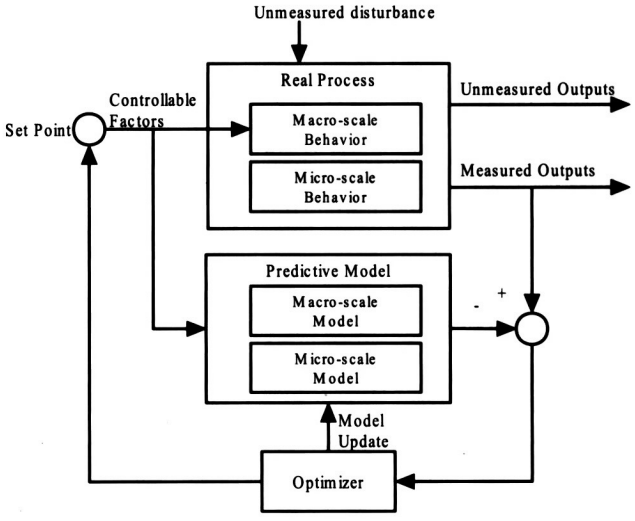


FIGURE 7. A Framework of Predictive Controllers with Multiscale Models.

can be employed in building the predictive model. In nano applications, a multiscale model can be used to characterize both macroscale and microscale behavior of a process. Usually, manipulated variables are applied to only the macroscale behavior of a process. Formulating an objective function is a key task in deriving the optimal set point. Most objectives in the multiscale control publications focused on the upper scale process performance. See Fenner et al. (2005) for a study of incorporating objectives from multiple process stages in deriving an automatic-control policy. The following gives another example.

The use of PI or predictive controllers has successfully improved process stability and product quality. However, in controlling a nanopositioner device, Salapaka et al. (2002) noted that the conventional PI controller does not meet the requirements for positioning. Instead, the authors considered an H-infinity controller to incorporate both performance and robustness in the objective function and found the optimum based on the H-infinity norm. As the objective is defined on multiple scales, the H-infinity norm is calculated as the supremum of the objective function on all scales. The optimization method based on H-infinity is a feasible way to solve multiscale problems and has substantially improved positioning speed and precision in the research.

Conclusion and Future Research

This article has reviewed applications of statistical techniques in nano technology. As the physical and chemical properties of nanoscale materials differ fundamentally from the properties of individual atoms and molecules or bulk matter, various challenges have occurred in designing, analyzing, and manufacturing nanodevices. Therefore, both conventional and new statistical techniques are expected to play important roles in nanotechnology research.

This paper first investigates the experimental design issues of nanodevice fabrication processes. Diverse application examples include fractional factorial designs, response surface designs, and robust parameter designs. For future nanotechnology research to make efficient use of these methods, we feel that the following issues are worthy of consideration:

(a) *Choice of experimental-design methods.* In some applications, simply full factorial or fractional factorial designs may be sufficient for model building and process understanding. However, as shown in several examples, nanofabrication processes can be very sensitive to small changes in controllable and noise factors. Moreover, the process outcomes may not be continuous random variables. In this case, irregular designs or new experimental design methods might be considered. Finally, special experimental constraints due to physical limitations may limit the use of certain designs.

(b) *Analysis of experimental data.* Due to the use of advanced measurement tools, spatial data, high-frequency signals, and/or qualitative measures are frequently observed in nano research. Combined with the multistage nature of the nanodevice–fabrication processes, statistical analysis of such complicated and/or large size data with many explanatory and noise factors provides many challenges. New algorithms, such as those developed by Jeong et al. (2006), Yuan and Kuo (2006), and Wang and Tsung (2007), might be useful for dealing with these new data types. Variable selection tools, such as Yuan et al. (2007), could be effective for dealing with large size factors constrained with experimental design structures. However, more research is needed in this field, especially for multistage, multiscale, and multilevel processes.

(c) *Use of computer simulations.* When the physical experimentation is costly and time consuming, the use of computers for process and device simulations is expected to become more prevalent in nano research. Sometimes, processes can be very complicated, leading to lengthy computer experiments. Thus, the design and analysis of computer experiments, as well as the integration between computer and physical experiments, are topics that deserve more future research efforts.

(d) *Interaction between statisticians and experts in material science, physics, and other disciplines.* Nanotechnology is a multidisciplinary subject. When statisticians are more involved in understanding the issues in nanoresearch, more effective procedures can be developed to deal with challenging application problems. Statisticians should be important players in novel scientific discoveries.

Quality issues have been emphasized in early production stages of nanodevices. The challenges confronted when applying SPC and APC techniques in nano applications are confounded with the new data types discussed above. Moreover, when online sensor information is available, it is interesting to see the of-

fline robust parameter designs extended (e.g., Joseph (2003)) to take advantage of the new information useful for process adjustments. See Edgar et al. (2000) and Del Castillo (2006) for a review in the topics of automatic control and statistical process adjustment.

Research on reliability is also critical to nanodevices and their fabrication processes. The intrinsic mechanism of failures for nanoproducts and the modeling of their lifetimes are areas of future research interest. Readers may refer to Jeng et al. (2007) for a detailed review in nanoreliability studies and see other papers in the same issue of the journal for specific topics. A different, but related, direction of new research is the "reliability" of measurement and positioning systems in nanomanufacturing. Because nanodevices are so small and the nanomanufacturing process requires speedy data collection, how can one know if the measurements are accurate and, more important, if the nanosubsystems are placed in the correct location within a very small-sized nanosystem? See Xia et al. (2007) and Qian and Wu (2007) for examples of new statistical methods developed for integrating information with different accuracy for statistical inference.

With the fast development of nanotechnology and the introduction of mass production of nanodevices, statistics is expected to play an increasingly important role in both academic and industrial fields. This review summarizes successful applications of statistical methods in nanotechnology seen in the literature, along with a few interesting topics for future research.

Acknowledgments

The authors would like to thank the editor and the anonymous referees for their insightful comments and suggestions, which have helped improve the quality of this review greatly. Dr. Jeng's research was partly supported by a grant from the Landmark Project of National Cheng Kung University. Dr. Wang's work was supported by the National Natural Science Foundation of China (NSFC) under grant 70802034. Dr. Lu's research was partially supported by the US National Science Foundation under the award 0400071.

References

General References

JENG, S.-L.; LU, J.-C.; and WANG, K. (2007). "A Review of Reliability Research on Nanotechnology". *IEEE Transactions on Reliability* 56(3), pp. 401–410.

NATIONAL NANOTECHNOLOGY INITIATIVE. (2008). "Nanotech Facts". Available at http://www.nano.gov/html/facts/home_facts.html.

ROCO, M. C. (2004). "Nanoscale Science and Engineering: Unifying and Transforming Tools". *AIChE Journal* 50(5), pp. 890–897.

Design of Experiments

BARGLIK-CHORY, C.; STROHM, C. R. H.; and MULLER, G. (2004). "Adjustment of the Band Gap Eenergies of Biostabilized CdS Nanoparticles by Aapplication of Statistical Design of Experiments". *Journal of Physical Chemistry, Series B* 108(23), pp. 7637–7640.

BASUMALLICK, A.; DAS, G. C.; and MUKHERJEE, S. (2003). "Design of Experiments for Synthesizing In Situ Ni-SiO2 and CO-SiO2 Nanocomposites by Non-isothermal Reduction Treatment". *Nanotechnology* 14(8), pp. 903–906.

COTA-SANCHEZ, G.; SOUCY, G.; HUCZKO, A.; and LANGE, H. (2005). "Induction Plasma Synthesis of Fullerenes and Nanotubes Using Carbon Black-Nickel Particles". *Carbon* 43(15), pp. 3153–3166.

DASGUPTA, T. (2007). "Robust Parameter Design for Automatically Controlled Ssystems and Nanostructure Synthesis". Ph.D. Dissertation, Georgia Institute of Technology, Atlanta, GA.

DASGUPTA, T.; MA, C.; JOSEPH, V. R.; WANG, Z. L.; and WU, C. F. J. (2008). "Statistical Modeling and Analysis for Robust Synthesis of Nanostructures". *Journal of the American Statistical Association* 103, pp. 594–603.

FASULO, P. D.; RODGERS, W. R.; OTTAVIANI, R. A.; and HUNTER, D. L. (2004). "Extrusion Processing of TPO Nanocomposites". *Polymer Engineering and Science* 44(6), pp. 1036–1045.

HOU, T.-H.; SU, C.-H.; and LIU, W.-L. (2007). "Parameters Optimization of a Nanoparticle Wet Milling Process Using the Taguchi Method, Response Surface Method and Genetic Algorithm". *Powder Technology* 173(3), pp. 153–162.

KIM, K. D.; HAN, D. N.; and KIM, H. T. (2004). "Optimization of Experimental Conditions Based on the Taguchi Robust Design for the Formation of Nano-Sized Silver Particles by Chemical Reduction Method". *Chemical Engineering Journal* 104(1–3), pp. 55–61.

KIM, K. D.; KIM, S. H.; and KIM, H. T. (2005). "Applying the Taguchi Method to the Optimization for the Synthesis of TiO2 Nanoparticles by Hydrolysis of TEOT in Micelles". *Colloids and Surfaces: A—Physicochemical and Engineering Aspects* 254(1–3), pp. 99–105.

LIN, T. Y.; DAVISON, K. L.; LEONG, W. S.; CHUA, S.; YAO, Y. F.; PAN, J. S.; CHAI, J. W.; TOH, K. C.; and TJIU, W. C. (2003). "Surface Topographical Characterization of Silver-Plated Film on the Wedge Bondability of Leaded IC Packages". *Microelectronics Reliability* 43, pp. 803–809.

MONTGOMERY, D. C. (2005). *Design and Analysis of Experiments*. New York, NY: Wiley.

NAZZAL, S.; NUTAN, M.; PALAMAKULA, A.; SHAH, R.; ZAGHLOUL, A. A.; and KHAN, M. A. (2002). "Optimization of a Self-Nanoemulsified Tablet Dosage Form of Ubiquinone Using Response Surface Methodology: Effect of Formulation Ingredients". *International Journal of Pharmaceutics* 240(1–2), pp. 103–114.

NEMBHARD, H. B.; ACHARYA, N.; AKTAN, M.; and KIM, S. (2006). "Design Issues and Analysis of Experiments in

Nanomanufacturing". In *Handbook of Industrial and Systems Engineering*, Badiru, A. B. ed., pp. 17.1–17.24. Boca Raton, FL: CRC Taylor & Francis.

PRAKOBVAITAYAKIT, M. and NIMMANNIT, U. (2003). "Optimization of Polylactic-Co-Glycolic Acid Nanoparticles Containing Itraconazole Using 2^3 Factorial Design". *AAPS PharmSciTech* 4(4), pp. 1–9.

TAGUCHI, G. (1986). *Introduction to Quality Engineering: Designing Quality into Products and Processes*. Tokyo: The Asian Productivity Organization.

YONG, V. and HAHN, H. T. (2005). "Dispersant Optimization Using Fesign of Rxperiments for SiC/Vinyl Ester Nanocomposites". *Nanotechnology* 16(4), pp. 354–360.

YÖRDEM, O. S.; PAPILA, M.; and MENCELOĞLU, Y. Z. (2008). "Effects of Electrospinning Parameters on Polyacrylonitrile Nanofiber Diameter: An Investigation by Response Surface Methodology". *Materials and Design* 29(1), pp. 34–44.

Data Collection, Analysis, and Modeling

BAZANT, Z. P. and PANG, S.-D. (2007). "Activation Energy Based Extreme Value Statistics and Size Effect in Brittle and Quasibrittle Fracture". *Journal of the Mechanics and Physics of Solids* 55, pp. 91–131.

BELICH, T. J.; SHEN, Z.; CAMPBELL, S. A.; and KAKALIOS, J. (2003). "Non-Gaussian $1/f$ Noise as a Probe of Long-Range Sstructural and Electronic Disorder in Amorphous Silicon". In *Proceedings of SPIE*, vol. 5112, pp. 67–77.

CHANG, I. J.; KANG, K.; MUKHOPADHYAY, S.; KIM, C. H.; and ROY, K. (2005). "Fast and Accurate Estimation of Nanoscaled SRAM Read Failure Probability Using Critical Point Sampling". In *Proceedings of the 2005 IEEE Custom Integrated Circuits Conference*, pp. 439–442.

CHEN, S. Y.; HUANG, Y. H.; and CAI, B. N. (2005). "Normal Distribution of Confinement Energy from a Photoluminescence Line Shape Aanalysis in Oxidized Porous Silicon". *Solid-State Electronics* 49(6), pp. 940–944.

CHEN, Y. P. and LEE, J. D. (2004). "Multiscale Modeling of Polycrystalline Silicon". *International Journal of Engineering Science* 42(10), pp. 987–1000.

GRANIER, J. J. and PANTOYA, M. L. (2004). "The Effect of Size Distribution on Burn Rate in Nanocomposite Thermites: A Probability Density Function Study". *Combustion Theory and Modeling* 8, pp. 555–565.

HSIEH, M.-S. and JENG, S.-L. (2007). "Accelerated Discrete Degradation Models for Leakage Current of Ultra-Thin Gate Oxides". *IEEE Transactions on Reliability* 56(3), pp. 369–380.

JEFFERSON, G.; GARMESTANI, H.; TANNENBAUM, R.; GOKHALE, A.; and TADD, E. (2005). "Two-Point Probability Distribution Function Aanalysis of Co-Polymer Nano-Composites". *International Journal of Plasticity* 21, pp. 185–198.

LI, X.; LE, J.; PILEGGI, L. T.; and STROJWAS, A. (2005). "Projection-Based Performance Modeling for Inter/Intra-Die Variations". In *Proceedings of the 2005 IEEE/ACM International Conference on Computer-Aided Design*, pp. 721–727.

MARGOLIN, G. and BARKAI, E. (2005). "Nonergodicity of Blinking Nanocrystals and Other Levy-Walk Processes". *Physical Review Letters* 94(8), 080601(1–4).

MIRANDA, E. and JIMENEZ, D. (2004). "A New Model for the Breakdown Dynamics of Ultra-Thin Gate Oxides Based on the Stochastic Logistic Differential Equation". In *Proceedings of the 24th International Conference on Microelectronics*, pp. 625–628.

ONODERA, H. (2006). "Variability: Modeling and Its Impact on Design". *IEICE Transactions on Electronics* E89-C(3), pp. 342–348.

QI, Y.; GAO, J.; and FORTES, J. (2005). "Markov Chains and Probabilistic Computation—A General Framework for Multiplexed Nanoelectronic Systems". *IEEE Transactions on Nanotechnology* 4(2), pp. 194–205.

REJIMON, T. and BAHNJA, S. (2005). "Time and Space Efficient Method for Accurate Computation of Error Detection Probabilities in VLSI Circuits". *Proceedings of the 2005 IEE Computers and Digital Techniques* 152(5), pp. 679–685.

SCHNEIDER, D.; SIEMROTH, P.; SCHULKE, T.; BERTHOLD, J.; SCHULTRICH, B.; SCHNEIDER, H. H.; OHR, R.; and PETEREIT, B. (2002). "Quality Control of Ultra-Thin and Super-Hard Coatings by Laser-Acoustics". *Surface and Coatings Technology* 153(2–3), pp. 252–260.

ZHAO, Y.; DEY, S.; and CHEN, L. (2004). "Double Sampling Data Checking Technique: An Online Testing Solution for Multisource Noise-Induced Errors on On-Chip Interconnects and Buses". *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 12(7), pp. 746–755.

Statistical Process Control and Quality Control

CHRISTOFIDES, P. D. and ARMAOU, A. (2006). "Control and Optimization of Multiscale Process Systems". *Computers & Chemical Engineering* 30(10–12), pp. 1670–1686.

DAS, T. K.; GANESAN, R.; SIKDER, A. K.; and KUMAR, A. (2005). "Online End Point Detection in CMP Using SPRT of Wavelet Decomposed Sensor Data". *IEEE Transactions on Semiconductor Manufacturing* 18(3), pp. 440–447.

FENNER, J. S.; JEONG, M. K.; and LU, J.-C. (2005), "Optimal Automatic Control of Multistage Production Processes". *IEEE Transactions on Semiconductor Manufacturing* 18(1), pp. 94–103.

GALLIVAN, M. A. (2005). "An Estimation Study for Control of a Lattice Model of Thin Film Deposition". *Computers & Chemical Engineering* 29(4), pp. 761–769.

GANESAN, R.; DAS, T. K.; SIKDER, A. K.; and KUMAR, A. (2003). "Wavelet-Based Identification of Delamination Defect in CMP (Cu-Low k) Using Nonstationary Acoustic Emission Signal". *IEEE Transactions on Semiconductor Manufacturing* 16(4), pp. 677–685.

JIANG, X. and BLUNT, L. (2004). "Third Generation Wavelet for the Extraction of Morphological Features from Micro and Nano Scalar Surfaces". *Wear* 257, pp. 1235–1240.

KLEPPER, C. C.; CARLSON, E. P.; HAZELTON, R. C.; YADLOWSKY, E. J.; FENG, B.; TAHER, M. A.; and MEYER, H. M. (2005). "H-Alpha Emission as a Feedback Control Sensor for Reactive Sputter Deposition of Nano-Structured, Diamond-Like Carbon Coatings". *IEEE Transactions on Plasma Science* 33(2), pp. 799–807.

LANTZ, M. A.; BINNIG, G. K.; DESPONT, M.; and DRECHSLER, U. (2005). "A Micromechanical Thermal Displacement Sensor with Nanometre Resolution". *Nanotechnology* 16(8), pp. 1089–1094.

LIN, T.; PAN, Y.; and HSIEH, C. (2003). "Precision-Limit Positioning of Direct Drive Systems with the Existence of Ffriction". *Control Engineering Practice* 11, pp. 233–244.

LOU, Y. and CHRISTOFIDES, P. D. (2003), "Estimation and Control of Surface Roughness in Thin Film Growth Using Kinetic Monte-Carlo Models". *Chemical Engineering Science* 58, pp. 3115–3129.

LOU, Y. M. and CHRISTOFIDES, P. D. (2006). "Nonlinear Feedback Control of Surface Roughness Using a Stochastic PDE: Design and Application to a Sputtering Process". *Industrial & Engineering Chemistry Research* 45(21), pp. 7177–7189.

MAE, K. and HONDA, T. (2000). "Growth Mode Variations of Thin Films on Nano-Faceted Substrates". *Thin Solid Films* 373(1–2), pp. 199–202.

NG, H. T.; LI, J.; SMITH, M. K.; NGUYEN, P.; CASSELL, A.; HAN, J.; and MEYYAPPAN, M. (2003). "Growth of Epitaxial Nanowires at the Junctions of Nanowalls". *Science* 300(5623), pp. 1249–1249.

NI, D. and CHRISTOFIDES, P. D. (2005a). "Dynamics and Control of Thin Ffilm Surface Microstructure in a Complex Deposition Process". *Chemical Engineering Science* 60(6), pp. 1603–1617.

NI, D. and CHRISTOFIDES, P. D. (2005b). "Multivariable Predictive Control of Thin Film Deposition Using a Stochastic PDE model". *Industrial & Engineering Chemistry Research* 44(8), pp. 2416–2427.

OHSHIMA, M. (2003). "Control and Design Problems in Material Processing—How Can Process Systems Engineers Contribute to Material Processing?" *Journal of Process Control* 13, pp. 599–605.

SALAPAKA, S.; SEBASTIAN, A.; CLEVELAND, J. P.; and SALAPAKA, M. V. (2002). "High Bandwidth Nano-Positioner: A Robust Control Approach". *Review of Scientific Instruments* 73(9), pp. 3232–3241.

SNEH, O.; CLARK-PHELPS, R. B.; LONDERGAN, A. R.; WINKLER, J.; and SEIDEL, T. E. (2002). "Thin Film Atomic Layer Deposition Equipment for Semiconductor Processing". *Thin Solid Films* 402(1–2), pp. 248–261.

VAN KAMPEN, N. G. (1992). *Stochastic Processes in Physics and Chemistry.* Amsterdam, The Netherlands: North-Holland.

Future Research

EDGAR, T. F.; BUTLER, S. W.; CAMPBELL, W. J.; PFEIFFER, C.; BODE, C.; HWANG, S. B.;, BALAKRISHNAN, K. S.; and HAHN, J. (2000). "Automatic Control in Microelectronics Manufacturing: Practices, Challenges, and Possibilities". *Automatica* 36(11), pp. 1567–1603.

DEL CASTILLO, E. (2006). "Statistical Process Adjustment: A Brief Retrospective, Current Status, and Some Opportunities for Further Work". *Statistica Neerlandica* 60(3), pp. 309–326.

JEONG, M. K.; LU, J.-C.; and WANG, N. (2006). "Wavelet-Based SPC Procedure for Complicated Functional Data". *International Journal of Production Research* 44(4), pp. 729–744.

QIAN, Z. and WU, C. F. J. (2007). "Bayesian Hierarchical Modeling for Integrating Low-Accuracy and High-Accuracy Experiments". *Technometrics* in press.

JOSEPH, V. R. (2003). "Robust Parameter Design with Feed-Forward Control". Technometrics 45, pp. 284–292.

SUNG, H. J.; SU, J.; BERGLUND, J. D.; RUSS, B. V.; MEREDITH, J. C.; and GALIS, Z. S. (2005). "The Use of Temperature-Composition Combinatorial Libraries to Study the Effects of Biodegradable Polymer Blend Surfaces on Vascular Cells". *Biomaterials* 26, pp. 4557–4567.

WANG, K. and TSUNG, F. (2007). "Run-to-Run Process Adjustment Using Categorical Observations". *Journal of Quality Technology* 39(4), pp. 312–325.

WANG, N.; LU, J.-C.; and KVAM, P. (2007). "Multi-Level Spatial Modeling and Decision-Making with Application in Logistics Systems". Technical report, The School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA.

XIA, H. F.; DING, Y.; and MALLICK, B. K. (2007). "Bayesian Hierarchical Model for Integrating Multi-Resolution Metrology Data". Paper presented in the best student paper competition, INFORMS Conference—Quality, Statistics and Reliability Section, Seattle, WA.

YUAN, M.; JOSEPH, V. R.; and LIN, Y. (2007). "An Efficient Variable Selection Approach for Analyzing Designed Experiments". *Technometrics* in press.

YUAN, T. and KUO, W. (2006). "Defect Pattern Recognition in Semiconductor Fabrication Using Model-Based Clustering and Bayesian Inference". Paper presented in the best student paper competition, INFORMS Conference—Quality, Statistics and Reliability Section, Pittsburgh, PA.

A Review of Reliability Research on Nanotechnology

Shuen-Lin Jeng, Jye-Chyi Lu, and Kaibo Wang

***Abstract*—Nano-reliability measures the ability of a nano-scaled product to perform its intended functionality. At the nano scale, the physical, chemical, and biological properties of materials differ in fundamental, valuable ways from the properties of individual atoms, molecules, or bulk matter. Conventional reliability theories need to be restudied to be applied to nano-engineering. Research on nano-reliability is extremely important due to the fact that nano-structure components account for a high proportion of costs, and serve critical roles in newly designed products. This review introduces the concepts of reliability to nano-technology; and presents the current work on identifying various physical failure mechanisms of nano-structured materials, and devices during fabrication process, and operation. Modeling techniques of degradation, reliability functions, and failure rates of nano-systems are also reviewed in this work.**

***Index Terms*—Degradation, failure analysis, nano-reliability.**

ACRONYM[1]

AFM	Atomic Force Microscope
CMOS	Complementary Metal-Oxide-Semiconductor
CNF	Carbon Nano Fiber
CO	Carbon Monoxide
CONAN	Configurable Nanostructures for Reliable Nano Electronics
CVD	Chemical Vapor Deposition
ECC	Error Correcting Codes
ESD	Electrostatic Discharge
FA	Failure Analysis
FT-IR	Fourier Transform Infrared
GC/MS	Gas Chromatography-Mass Spectroscopy
GPC	Gel Permeation Chromatography
HCI	Hot Carrier Injection
IC	Integrated Circuit
ITRS	International Technology Roadmap for Semiconductors
MEMS	Micro-Electro-Mechanical Systems
MIM	Metal-Insulator-Metal
MOS	Metal-Oxide-Semiconductor
MOSFET	Metal-Oxide-Semiconductor Field-Effect Transistor
MRAM	Magnetoresistive Random Access Memory
NAND	Negated Conjunction
nanoDAC	Nano Deformation Analysis by Correlation
NBTI	Negative Bias Temperature Instability
NDE	Nondestructive Evaluation
NEMS	Nano-Electro-Mechanical Systems
NF	Nano-Fibrous
nMOS	Negative-Channel Metal-Oxide-Semiconductor
NSET	The Nanoscale Science, Engineering, and Technology
PAS	Positron Annihilation Spectroscopy
PLLA	Poly-L-Lactide Acid
PMA	Post Metal Annealing
PRISM	Probabilistic Symbolic Model Checker
PS	Polystyrene
RF	Radio-Frequency
SC-Si	Single Crystal Silicon
SEM	Scanning Electron Microscopy
SIA	Semiconductor Industry Association
SOC	System-on-a-Chip
SPC	Statistical Process Control
SPM	Scanning Probe Microscopy
SRAM	Static Random Access Memory
SW	Solid-Walled
TDDB	Time-Dependent Dielectric Breakdown
UNCD	Ultra-Nano-Crystalline-Diamond
UV	Ultraviolet
WS2	Tungsten Disulfide

Manuscript received January 8, 2007; revised February 21, 2007; accepted March 8, 2007. Associate Editor: S. Bae.

S.-L. Jeng is with the Department of Statistics, National Cheng Kung University, Tainan 701, Taiwan (e-mail: sljeng@mail.ncku.edu.tw).

J.-C. Lu is with the School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: jclu@isye.gatech.edu).

K. Wang is with the Department of Industrial Engineering, Tsinghua University, Beijing 100084, China (e-mail: kbwang@tsinghua.edu.cn).

Digital Object Identifier 10.1109/TR.2007.903188

[1]The singular and plural of an acronym are always spelled the same.

I. INTRODUCTION

THE ability to measure, and manipulate matter at the atomic/molecular scale has led to the discovery of novel materials. A nanometer is 10^{-12} meter; a single human hair is about 8×10^5 nanometers wide.

According to the definition of The Nanoscale Science, Engineering, and Technology (NSET) Subcommittee of the National Science and Technology Council's Committee on Technology (Roco [63]), "Nanotechnology is the research, and technology development at the atomic, molecular, or macromolecular levels, in the length scale of approximately 1–100 nanometer range, to provide a fundamental understanding of phenomena, and materials at the nanoscale; and to create, and use structures, devices, and systems that have novel properties, and functions

because of their small, and/or intermediate size." At this level, the physical, chemical, and biological properties of materials differ in fundamental, valuable ways from the properties of individual atoms, molecules, or bulk matter.

Over the last ten years, there has been a series of critical convergences in previously disparate technology areas. Integration of IC within complex MEMS device structures have opened vast new avenues for integrated sensor, and system technologies. Likewise, integration of functional "smart" nanomaterials into self-standing MEMS devices has dramatically widened the functional application space for miniaturized systems including microfluidic, and bio-compatible microdevices for biomedical applications.

As a consequence, the next several decades will see unprecedented levels of integration of emerging nanomaterials, nanoelectronic architectures, and nano-MEMS platforms. This will pose severe challenges for testing, reliability, and metrology techniques required to support such development. The integration of varied material, and technology approaches has, and will continue to result in the combination of heretofore field-specific testing, reliability, and metrology methodologies. For example, the adaptation of tomographic approaches pioneered in the health, and nondestructive evaluation (NDE) analysis fields to destructive focused ion beam imaging has generated significant interest for nanoscale 3D reconstruction in IC, and NEMS metrology. This type of "cross-contamination" between the traditional fields of NDE/testing/reliability, and the emerging areas of nanomaterials, nanoelectronics, and NEMS is essential to develop the metrology, and test/reliability solutions that are needed. Near-field acoustics for nanoscale mechanics & stress evaluation, near-field optics for nanoscale chemical & optical probing, and nanometer-resolved x-ray imaging are additional examples of such "cross-contamination".

Engineers are needed to help increase reliability, while maintaining effective production schedules to produce current, and future electronics at the lowest possible cost. Without effective quality control, devices dependent on nanotechnology will experience high manufacturing costs, including transistors which could result in a disruption of the continually steady Moore's law. Nanotechnology can potentially transform civilization. Realization of this potential needs a fundamental understanding of friction at the atomic scale. Furthermore, the tribological considerations of these systems are expected to be an integral aspect of the system design, and will depend on the training of both existing, and future scientists, and engineers in the nano scale (Krim [37]).

Nano-reliability measures the probability that a nano-scaled product performs its intended functionality without failure under given conditions for a specified period of time. Experience in conventional manufacturing shows that neglecting reliability in an early stage results in extremely high direct, and indirect costs on production suspension, product repair or replacement, and other loss in later stages of product lifecycles. In macro- and micro-systems, reliability has also been essential for product design, and manufacturing (see, e.g., Srikar & Senturia [70], De Wolf [18], van Spengen [77], Melle, *et al.* [50], and Choa [16]). In the promising nano-world, reliability will be even more important due to expected higher functionality, and complexity of products. As most materials exhibit totally different physical properties when operating in a nano scale, compared to larger scales, research on nano-reliability defines a new, promising area that deserves more interests.

In nano-reliability research, traditional generic reliability theories could still be applicable with proper modifications. However, new models & theories are also needed to characterize diverse behaviors that happen in the nano-world. Physical processes do not scale with size, and time. When size goes to a scale as small as micro or nano, dramatic changes in electrical conductivity, reaction kinetics, corrosion processes, etc. may be seen. Important concepts in reliability engineering, such as fatigue, friction, damping, wear-out, and repair mechanisms, have different physical meanings on atomic or molecular scales.

The reliability of nano devices is still far from perfected. Failures in micro-systems, and nano-systems can be traced back to thermal, mechanical, chemical, electrical, or combined origins thereof; which may be caused by different manufacturing stages such as wafer processing, packaging, and final assembly; and post-production stages such as transportation, and usage. Typical failures found in these systems are cracks, delamination, buckling, warpage, popcorning, stress voiding, fatigue, pattern shift, thermo-migration, and electrical stress-induced failures such as hot carrier degradation, breakdown of thin oxides, and electro-migration. The majority of these failures (65%) are thermo-mechanically related (Michel [51]).

This review coves extensively the latest progress on nano-reliability published in the following journals: Microelectronics Reliability, Microelectronic Engineering, IEEE TRANSACTIONS ON DEVICE AND MATERIALS RELIABILITY, IEEE TRANSACTIONS ON NANOTECHNOLOGY, Proceedings of SPIE, IEEE TRANSACTIONS ON ELECTRON DEVICES, IIE Transactions and Materials Science and Engineering, and others.

As an emerging field, research on nano-reliability is facing the following main tasks:

- Introduction of concepts, and technical terms of reliability to nano-technology in an early state.
- Identification of physical failure mechanisms of nano-structured materials, and devices during fabrication process and operation.
- Determination of quality parameters of nano-devices, failure modes, and failure analysis including reliability testing procedures, and instrumentation to localize nano-defects.
- Modeling of reliability functions, and failure rates of nano-systems.

The rest of this paper is organized as follows. Due to the promising development of nanotechnology, review work that focuses on different aspects of nanotechnology appear. Section II introduces several other review papers that are related to nano-reliability research. In reliability research, it is important to understand the types of failure-modes, and intrinsic mechanism, so Section III reviews failure-modes commonly seen in high-k films, nanostructure, MEMS, CMOS, and other nano-structured components. Manufacturing reliability, aging, and degradation models as well as failure, and lifetime models are also reviewed in this section. Section IV investigates reliability testing issues, and related evaluation & measurement techniques. Section V in-

troduces some structure, and parameter design methodologies for nano-reliability. Finally, Section VI concludes this paper with a summary, and view of future research.

II. Related Review Work

As nanotechnology, and MEMS are enabling new discoveries in diverse fields of science, and engineering (Huff [25]), a collection of review papers have been seen recently that try to summarize the various impacts that nanotechnology is bringing. Focusing on different physical, and statistical issues, these review papers have provided a general background of reliability research in nano-engineering.

Motivated by a recent prediction made by the Semiconductor Industry Association (SIA) in the International Technology Roadmap for Semiconductors (ITRS [26]) that the silicon technology will continue its historical rate of advancement with the Moore's law for at least a couple of decades, Wong & Iwai [78] indicated that the silicon gate oxide will be scaled down to its physical limit. An alternate way is to replace oxide with a physically thicker high-k material to help solve most of the problems. However, new problems concerning reliability, and performance have to be addressed. Ribes, *et al.* [61] reviewed the status of reliability studies of high-k gate dielectrics, and illustrated some concepts with experimental results.

Stathis [71] focused on the reliability limits for the gate insulator in CMOS technology. The author reported that present research is aimed at better understanding the nature of the electrical conduction through a breakdown spot, and the effect of the oxide breakdown on device, and circuit performance. However, it is also noted that an oxide breakdown may not necessarily lead to immediate circuit failure. Therefore, more research is needed to develop a quantitative methodology for predicting the reliability of circuits. Lombardo, *et al.* [45] reviewed the subject of oxide breakdown, while focused more on the case of the gate dielectrics of interest for situations under which Si oxides or oxynitrides of thickness ranging from some tens of nanometers down to about 1 nm. Specifically, the authors investigated the kinetics of oxide degradation, and the statistics of the time to breakdown. Experimental studies were conducted to study the influential factors to oxide breakdown.

The only commercialized electronic packaging technique is still lead-bearing soldering. Challenging issues, such as lower electrical conductivity, conductivity fatigue in reliability testing, limited current-carrying capability, and poor impact strength, are not well solved by other new techniques. Li & Wong [42] gave a thorough review of the recent development in electrical conductive adhesives, and studied the electrical, mechanical, and thermal behavior improvements, as well as reliability enhancement under various conditions.

Microstructural design has attracted increasing interest in the modern development of hard coatings for wear-resistant applications. Mayrhofer, *et al.* [49] demonstrated the correlation between microstructure, mechanical properties, and tribological properties of hard ceramic coatings. The authors also noted that developments in all applications will benefit from a closer interaction of the different fields as many of the materials quality & reliability issues are similar, for example, for controller file texture, defect density, and purity.

The development of the nanotechnology has extended its impact to the degradation of solid dielectrics as well. Morshuis [53] reviewed the vast literature on partial discharge, and partial discharge induced degradation. The author emphasized that many properties of the new generation of dielectrics can be affected by the introduction of small amounts of nano-sized particles.

Electrostatic discharge (ESD) protection design for mixed-voltage I/O interfaces has been one of the key challenges of system-on-a-chip (SOC) implementation in nano-scaled CMOS processes. Ker & Lin [29] presented a broad overview on the ESD design constraints in mixed-voltage I/O interfaces, the classification & analysis of ESD protection designs for mixed-voltage I/O interfaces, and the designs of high-voltage-tolerant power-rail ESD clamp circuits.

As nanotechnology is gradually being integrated in new product design, it is important to understand the mechanical, and material properties for the sake of both scientific interest, and engineering usefulness. Kitamura, *et al.* [32] reviewed works on the strength of ideal nano-structure components. The authors also noted that a good understanding of the mechanical properties of nano-structured components is important to the design of fabrication/assembly processes, and reliability in service, which will soon be a major focus when nanotechnology is ready for massive production. Some discussion on thermally driven reliability issues in microelectronic systems can be found in the work of Lasance [39].

III. Mechanism Analysis and Modeling in Nano-Reliability

As aforementioned, failures in micro-systems, and nano-systems can be traced back to thermal, mechanical, chemical, electrical, or combined origins thereof. In this section, we review the physical, mechanical, or chemical failure modes of popular applications in the current nano-engineering research, and introduce research in the modeling of these failures.

A. Failure-Mode and Mechanism Analysis

The behavior of nanostructured materials/small-volume structures, and biological/bio-implantable materials is currently much in vogue in materials science. One aspect of this field, which has received only limited attention, is their fracture & fatigue properties. Ritchie, *et al.* [62] examined the premature fatigue failure of silicon-based micron-scale structures for MEMS, and the fracture properties of mineralized tissue, specifically human bone.

Similarly, accurate identification of mechanical properties arises whenever very thin coatings that consist of single or multiple layers are considered. Rapid developments in the areas of nano-fabrication, nano-manipulation, and nanotechnology lead to the increased importance of reliable characterization of mechanical properties of progressively thinner coatings. A recent study on this topic is Korsunsky & Constantinescui [34].

Failure analysis (FA) also plays an important role in the development, and manufacture of integrated circuits. However, instrumental limits are already causing problems in FA in the tenth-micron CMOS realm. Nanoelectronic devices will meet the problem of incapable analytical tools. Vallett [76] introduced

state-of-the-art microelectronic failure analysis processes, instrumentation, and principles. The major limitations, and future prospects determined from industry roadmaps are discussed. Specifically highlighted is the need for a fault isolation methodology for failure analysis of fully integrated nanoelectronics devices.

Nanocomposites exhibit new, improved properties when compared to their micro- or macrocomposite counterparts. By lowering the particle size to nano dimensions, the special effects in polymer composites appear. Kovacevic, *et al.* [36] compared the properties of composites with micro-, and nano-sized calcium carbonate (CaCO3) particles in a poly (vinyl acetate) (PVAc) matrix. Mathematical models were used to quantify the interfacial interactions in the composites under investigation. It seems that a key characteristic of the nanocomposites is the formation of a three-dimensional interphase with a significant amount of restricted chain mobility.

Luo, *et al.* [46] examined some fundamental reliability aspects of high-k film through ramp voltage stress testing. By studying dielectric relaxation, and analysing the TRANSIENT conductivity, breakdown modes of the tested high-k film are identified; a sensitive method of breakdown detection in ramped voltage tests is proposed, and investigated.

Luo, *et al.* [47] investigated the breakdown phenomena of TiN/ZrHfO/HfSiO/p-Si, and found that defect accumulation in the interface region triggers breakdown of the stack subjected to gate injection. Luo, *et al.* [48] investigated the relaxation currents of a variety of high-k gate stacks, and obtained evidence relating relaxation properties to dielectric integrity. The authors suggested that even though relaxation current is not detectable on SiO_2, it is obvious on the high-k stacks, which signified the integrity of high-k dielectrics. The breakdown sequence of individual layers in the double-layer high-k stacks has been identified. Such findings would be valuable in understanding the breakdown of multilayer high-k stacks.

Lombardo, *et al.* [45] presented new failure mechanisms associated with breakdown in high-k gate stacks on the case of Si oxides, or oxynitrides of thickness ranging from some tens of nanometers down to about 1 nm. In addition to dielectric-breakdown-induced epitaxy commonly found in breakdowns in poly-Si/SiON, and poly-Si/Si_3N_4 MOSFETs, grain-boundary, and field-assisted breakdowns near the poly-Si edge are found. The authors also developed a model based on breakdown induced thermo-chemical reactions to describe the physical microstructural damages triggered by breakdown in the high-k gate stack, and the associated post-breakdown electrical performance.

Bae, *et al.* [2] provided basic physical modeling for MOSFET devices based on the nano-level degradation that takes place at defect sites in the MOSFET gate oxide. The authors investigated the distribution of hot-electron activation energies, and derived a logistic mixture distribution using physical principles on the nanoscale.

Basaran, *et al.* [4] qualified the damage mechanism in solder joints in electronic packaging under thermal fatigue loading through experiments. The authors also showed that damage MECHANISMS under thermal cycling are very different than those under mechanical cycling. Elastic modulus degradation under thermal cycling, which is considered as a physically detectable quantity of material degradation, was measured using a nano-indenter.

Tomczak, *et al.* [74] presented the use of single molecules to study local, and nanoscale polymer dynamics. Fluorescence lifetime fluctuations were used to extract the number of polymer segments taking part in the rearranging volume around the probe molecule below the glass transition temperature. It was found that the number of polymer segments decreased with increasing temperature.

B. Manufacturing Reliability

The development of nanotechnology will lead to the introduction of new products to the public. In the modern large-scale manufacturing era, reliability issues have to be studied; and results incorporated into the design, and manufacturing phases of new products.

Kitamura, *et al.* [32] pointed out that some nano structured components, including nano films, nano tubes, and nano clusters, are very reliable once manufactured. Such components show high strength characteristics. However, their manufacturing procedures are complicated. Moreover, utilizing the structure at the nano level is a key technology in the development of electronic devices, and elements of nano electro-mechanical systems. Therefore, it is important to understand the mechanical properties for engineering usefulness, such as design of fabrication processes, to produce these components effectively. Lee, *et al.* [40] discussed the critical issues of MEMS in four categories: functional interfaces, reliability, modeling, and integration. They conducted burn-ins, and accelerated tests to ensure the production of a reliable MEMS device. In the nanofabrication of solid materials, Klein-Wiele, *et al.* [33] found that there is a quality & reliability advantage of the combination of femtosecond pulse durations with ultraviolet wavelengths in the nanofabrication of solid materials.

Kouvelis & Mukhopadhyay [35] analyzed failures, such as access time failure, read/write stability failure, and hold stability failure in the stand-by mode of SRAM cells due to process parameter variations; and they modeled the failure probabilities. A method to predict the yield of a SRAM memory chip designed with a cell is proposed based on the cell failure probability. The developed method can be used in the early stage of a design cycle to optimize the design for yield enhancement.

As one of the critical procedures in semiconductor manufacturing, nanoimprint lithography is seen as an alternative to conventional nanometer scale patterning technologies like electron beam lithography. Finder, *et al.* [19] introduced that nanoimprint lithography is a low cost method for the fabrication of nano-scaled patterns. The parallel process involves a stamp pressing into a softened polymer layer. This method has demonstrated high reliability, high throughput, and low cost. It has been used to print on 6 inches wafers, and has demonstrated the ability to master nanostructures down to 10 nm.

On-chip integrated MIM capacitors are finding increasing attention for various applications in advanced high performance mixed signal, and RF products. Typical requirements include low area consumption, large specific capacitance, low capacitance tolerances, high quality factors, and low parasitic substrate

coupling. Schrenk, *et al.* [66] presented an approach for integrating MIM caps into a copper multilevel metallization using Cu lines as a bottom electrode for the capacitor.

Sikora, *et al.* [68] examined the various technologies for implementing embedded flash cells. By focusing on automotive application, the authors discussed the technological basics with regard to practical consequences, and concentrated on the aspects of reliability of embedded flash cells. The required process steps are presented as well as the test approaches to ensure a high-quality production level.

Ganesan, *et al.* [20] utilized multi-scale wavelet SPC to analyze the quality of chemical mechanical planarization of silicon wafers in production. The wavelet method allows for real-time defect detection during manufacturing on the nano scale. By integrating with an advanced SPC method, individual defects could be differentiated simultaneously as they occur in the chemical mechanical planarization processing.

C. Aging and Degradation Models

The aging, and degradation effects are the major reasons that lead to product failures. In this section, research on photodegradation, oxide breakdown, and other aging & degradation models are reviewed.

Sivalingam & Madras [69] analyzed the mechanism of photodegradation. Product degradation was detected using UV-Visible spectroscopy, FT-IR, and GC/MS. The mechanism of breakage during photocatalytic degradation can be attributed to concerted rearrangement (Photo-Fries), and non-concerted cage recombination (oxidation). The degradation was measured by the molecular weight distribution using gel permeation chromatography (GPC), and modeled with continuous kinetics distribution.

Liu, *et al.* [44] developed a model accounting for oxide breakdown. Data measured on the nMOS transistor biased in the linear region before, and after breakdown are used to extract the breakdown spot resistance, and total gate capacitance. This methodology provides critical information about the impact brought by gate oxide breakdown.

Lin, *et al.* [43] proposed a new sub-circuit degradation model. The reliability of class-E, and class-A power amplifiers is investigated. Experimental results of degradation characteristics on the fabricated circuits agree well with the simulation predictions. From this newly developed model, the authors found that the class-E amplifier degrades faster than a class-A amplifier. A shorter lifetime is expected for a class-E amplifier.

Umemura, *et al.* [75] clarified the degradation behavior of the electric double layer capacitors. The authors identified that the capacitance decrease might be caused by the degradation of the electrolysis, propylene carbonate, and also the degradation by-products which piled on the nano-scaled micro-cavity surface of the activated-carbon particles of the electrodes. The degradation mechanism was found to be governed by the Arrhenius chemical kinetics theory, and consisted of two stages with different activation energies.

Cester, *et al.* [10], and Miranda & Jimenez [52] investigated the breakdown dynamics of ultrathin SiO_2 films in metal-oxide-semiconductor structures. It was shown that the progressive increase of the leakage current that flows through the oxide when subjected to constant electrical stress can be modeled by the stochastic logistic differential equation. This approach relies on a time-scale separation in which a deterministic term provides the S-shaped growth trajectory, while a second term of the equation deals with the noisy behavior. Because of the inherent mean reverting property of the simulation process, the proposed model is also able to cover cases in which sudden upward, and downward changes of the system's conductance are registered.

By combining the oxide time to breakdown model with a defect size distribution, Kim, *et al.* [30] presented a model to tie oxide yield to time-dependent reliability. Cester, *et al.* [10] presented an original model to explain the accelerated wear-out behavior of irradiated ultra-thin oxides. The model uses a statistical approach to model the breakdown occurrences based on a non-homogeneous Poisson process.

Suehle, *et al.* [73] studied two post soft-breakdown modes: one in which the conducting filament is stable until hard breakdown, and one in which the filament continually degrades with time. Acceleration factors are different for each mode, indicating different physical mechanisms. The results suggest that the "hardness" of the first breakdown influences the residual time distribution of the following hard breakdown. Tunneling current appears to be the driving force for both modes.

In practice, degradation processes do not all occur in a continuous pattern. Hsieh & Jeng [24] established a procedure for accessing the reliability using a discrete model. A non-homogeneous Weibull compound Poisson model with accelerated stress variables is considered by the authors. A dataset measuring the leakage current of nanometer-scale gate oxides is analyzed by using this procedure.

Chen & Ma [14] examined in-vitro hydrolytic degradation behavior for nano-fibrous (NF) poly (L-lactic acid) (PLLA) foams prepared by phase separation. In their research, nano-fibrous foams were incubated in phosphate-buffered saline at 371° C for 15 months. Upon removal, changes in mass, molar mass, morphology, BET specific surface area, mechanical properties, and thermal properties were compared with those of similarly incubated solid-walled (SW) PLLA foams. The initial surface area in NF foams was over 80 times higher than in SW foams. During incubation, NF surface area decreased steadily, only possessing 17% of the original specific surface area after 15 months, and SW surface area stayed constant throughout.

Polymer additives often show many side reactions during the aging of polymers, which changes the useful lifetime of materials. Side reactions have been found in nano-grade titanium dioxide additives. The interaction between titanium dioxide pigments, and stabilizers therefore is proposed as a field of great importance. Zeynalov & Allen [79] investigated the influence of nano, and micron particle grade anatase; and rutile titanium dioxide pigments on the efficiency of a hindered amine stabilizer. Bressers, *et al.* [9] presented some results of models to link the chemical details of polymers, and microelectronic reliability.

Kufluoglu & Alam [38] investigated Negative Bias Temperature Instability (NBTI)- induced degradation for ultra-scaled, and future-generation MOSFET. Numerical simulations based on Reaction-Diffusion framework were implemented. Geometric dependence of degradation arising from the transistor

structure & scaling is incorporated into the model. The simulations are applied to narrow-width planar triple-gate, and surround-gate MOSFET geometries to estimate the NBTI reliability under several scaling scenarios. Unless the operating voltages are optimized for specific geometry of transistor cross section, the results imply worsened NBTI reliability for the future-generation devices based on the geometric interpretation of the NBTI degradation. A time-based model is developed to predict the degradation.

D. Failure and Lifetime Models

Failures, and lifetime models are important to lifetime forecasting, and design for reliability. Research on such models for CMOS, digital micromirror devices, nanotubes, and other important nano-devices are recently seen in the literature.

Schwalke, *et al.* [67] investigated the breakdown of extra thick gate oxides (50–150 nm) used in power MOS devise. Weibull probability plots are used to describe the failure distribution of the thick gate oxides.

Groeseneken, *et al.* [21] reviewed an acceleration model. Recent trends of reliability assessment in CMOS were also discussed. Lee, *et al.* [41] investigated the possibility of integrating chemical vapor deposition (CVD) HfO_2 into the multiple gate dielectric SOC process in the range of 6–7 nm. They predicted the ten-year time-dependent dielectric breakdown (TDDB) reliability of HfO/SiO_2 gate stack.

Namazu & Isono [54] studied the effects of specimen size, frequency, and temperature on fatigue lives of nanoscale single crystal silicon (SC-Si), and silicon dioxide (SiO_2) wires for reliable design of micro/nano electromechanical systems. Evaluation of fatigue lives for nanoscale fixed SC-Si, and Si02 wires was performed by stress-controlled cyclic bending tests under an atomic force microscope (AFM) at temperatures ranging from 295 K to 573 K. In MEMS-00, and MEMS-01, the quasi-static bending tests under the AFM for nanoscale SC-Si wires were reported.

Barber, *et al.* [3] summarized, and discussed the limited statistically significant, currently available, experimental data for the tensile strength of individual nanotubes of any sort. Only three such data sets currently exist: two for multi-wall carbon nanotubes, and one for multi-wall WS2 nanotubes. It is shown by the authors that Weibull-Poisson statistics accurately fit all strength data sets, and thus seem to apply at the nano scale as well as at the micro/macro-scales.

Jean, *et al.* [27] used positron annihilation spectroscopy (PAS), coupled with a variable mono-energetic positron beam, to investigate surface, and interfacial properties in thin polymeric films. The authors measured free-volume properties from ortho-Positronium lifetime, and the S parameter of Doppler broadening of energy spectra from annihilation radiation as a function of the depth and temperature in thin polymeric films. They also presented glass transition temperature profile on nanoscale layered structures in polystyrene (PS) thin films.

Reliability of magnetic tunnel junctions emerges as a critical problem for the successful application of the new writing schemes to the next generation high-density MRAM devices. Kim, *et al.* [31] presented reliability characteristics, and the thermal stability of magnetic tunnel junctions. The Weibull distribution is used for the data fitting.

Chasiotis & McCarty [11] described strength data for uniformly stressed MEMS specimens. The Weibull model has been used extensively. The authors investigated the relevance of the Weibull model to more general situations of MEMS-scale specimen failure. The applicability of the Weibull model for describing the failure of (a) specimens with a single flaw distribution, and variable geometry; and (b) multiple flaw populations, and a specific geometry is studied.

IV. Reliability Testing and Evaluation

Reliability testing of nano-devices is important in nano-reliability research. This section reviews techniques used for films, wafers, nanocomposite materials, and other nano-products.

A. Reliability Testing

Thin films, and coatings can fail by fatigue at lower loads than predicted by static tests. Beake & Smith [5] introduced a nano-impact technique, which is a low load impact test capable of revealing remarkable differences in performance useful in optimizing the design of coating systems for improved durability. In particular, the proposed technique can be used to find the optimum coating process parameters for enhanced toughness, and damage tolerance; and also differentiate between cohesive (chipping), and adhesive failure (delamination).

Chen, *et al.* [13] investigated the reliability of anodically-bonded packages between silicon, and borosilicate glass wafers. Effects of certain accelerated environmental tests such as thermal cycling, thermal shock, and boiling test on bonding quality are evaluated. Bonding strength is measured using an in-house tensile tester. The fact that fracture mainly occurs inside the glass wafer rather than along the interface indicates the robustness of the bond.

Pervin, *et al.* [58] developed a novel technique to fabricate nano-composite materials containing SC-15 epoxy resin, and carbon nano fiber (CNF). A high-intensity ultrasonic liquid processor was used to obtain a homogeneous molecular mixture of epoxy resin, and carbon nano fiber. Based on the experimental results, a nonlinear damage model was established to describe the stress-strain relationship of the epoxy, and its nano-composite.

Cheung [15] noted that solid-insulator breakdown always leads to an irreversible permanent conduction path. This is a key assumption in all gate-oxide breakdown reliability assessment, and lifetime projection. This assumption is not valid when the gate-oxide thickness is less than 2 nm, and the operation voltage is 1 V or less. Chatterjee, *et al.* [12] investigated the dielectric breakdown property of ultrathin 2.5, and 5.0 nm hafnium oxide (HfO_2) gate dielectric layers with metal nitride (TaN) gate electrodes for MOS structure. Reliability studies were performed with constant voltage stress to verify the effects of changing processing conditions (film thicknesses, and post metal annealing temperatures) on time to breakdown. The leakage current characteristics are improved with post metal annealing (PMA) temperatures for both 2.5, and 5.0 nm thicknesses HfO_2.

B. Reliability Evaluation and Measurement Techniques

Measurement, and evaluation of reliability of nano-devices is an important subject. New technology is developed to support the achievement of this task.

As noted by Keller, *et al.* [28], with ongoing miniaturization from MEMS towards NEMS, there is a need for new reliability concepts making use of meso-type (micro to nano) or fully nano-mechanical approaches. Experimental verification will be the major method for understanding theoretical models, and simulation tools. Therefore, there is a need for developing measurement techniques which have capabilities of evaluating strain fields with very local (nanoscale) resolution.

Peng & Cho [57] proposed the concept of a new type of nanoscale sensor devices that can detect the presence of CO, and water molecules. To overcome the reliability problem, these devices were developed by substitution-doping of impurity atoms (such as boron, or nitrogen atoms) into intrinsic single-wall carbon nanotubes, or by using composite nanotubes. Keller, *et al.* [28] developed the nanoDAC method (nano Deformation Analysis by Correlation), which enables the extraction of nanoscale displacement fields from scanning probe microscopy (SPM) images.

Su, *et al.* [72] used the Pearson correlation coefficient to calculate the digital speckle correlation for nano-metrology, which can be applied to MEMS, and IC packaging. Bhaduri & Shukla [7] used a Markov Random Field as a model of computation for nanoscale logic gates. They take the approach further by automating this computational scheme using a Belief Propagation algorithm (Pearl [56]).

Bhaduri & Shukla [8], and Bhaduri & Shukla [6] extended previous work on evaluating reliability-redundancy trade-offs for NAND multiplexing to trade-offs among granularity, redundancy, and reliability for several redundancy mechanisms; and presented their automation mechanism using the probabilistic model checking tool PRISM. Nano computing in the form of quantum, molecular, and other computing models is proliferating as nano fabrication advances. With the advance to nano scale fabrication, unprecedented levels of defects arise. Their MATLAB-based tool NANOLAB helps to uncover the anomalies during fabrication, thereby providing better insight into defect tolerant design decisions.

Norman, *et al.* [55] evaluated the reliability of defect-tolerant architectures for nanotechnology with probabilistic model checking.

Holmberg [23] dealt with the role of tribology in the large, complex scope of reliability engineering. They discussed different tribology-related methods for improving product reliability, such as reliability design, component lifetime, condition monitoring, and diagnostics. Cumulative wear, and change of friction were recorded through time.

Zhihong, *et al.* [81] presented the methodology of the reliability modeling, and simulation for the state-of-the-art nanotechnology; and discussed the extraction for HCI (Hot Carrier Injection), and NBTI (Negative Bias Temperature Instability) for product lifetime models. The integration of these models into the transistor level, and gate level simulation flow can be used by the designers to satisfy product reliability requirements.

V. Structure and Parameter Design for Reliability

Due to the high reliability-related cost, structure & parameter design techniques can be employed to minimize possible loss due to poor quality nano-devices. The following work provides useful examples of design-for-reliability in nano-engineering.

Zhao, *et al.* [80] presented a "Noise Impact Analysis" methodology to evaluate the transient error effects in static CMOS digital circuits. A "Noise Capture Raito" has been defined by the authors to measure the transient noise effects in the circuit. The proposed methodology facilitates the economic design of robust nanometer circuit.

Prabhakumar, *et al.* [59] described the assembly, and reliability of flip chips with a nano-filled wafer. They use Box-plots to compare the underfill interface distributions, and conducted a reliability test for failure mechanisms.

Nanotechnology in static random access memory (SRAM) cells is also advancing. SRAM is much more reliable than dynamic RAM, and allows for access time to be much quicker. Agarwal, *et al.* [1] conducted experimental analysis of the impact of process variation on the different failure mechanisms in SRAM cells on a sixty-four K cache. The authors proposed a process tolerant cache architecture, which can achieve ninety-four percent yield compared to its original thirty-three percent yield (standard cache) in the forty-five nanometer predictive technology. This technique surpasses all the contemporary fault tolerant schemes such as row-column redundancy, and ECC in handling failures due to process variation.

Pugno, *et al.* [60] used an experimental-theoretical method to investigate the strength of structures having complex geometries, which is commonly used in MEMS. It involves the stretching to failure of freestanding thin film membranes, in a fixed configuration, containing micro fabricated sharp cracks, blunt notches, and re-entrant corners. The defects, made by nano-indentation, and focused ion beam milling, are characterized by scanning electron microscopy (SEM). MEMS structures made of ultra-nano-crystalline-diamond (UNCD) were investigated using this methodology. A theory to predict the strength of micro structures with defects was proposed, and compared with experimental results. It was shown that the concepts of fracture mechanics can be applied with confidence in the design of MEMS.

Schmid & Leblebici [65] discussed various circuit-, and system-level design challenges for nanometer-scale devices, and single-electron transistors, with an emphasis on the functional robustness, and fault tolerance point of view. A set of general guidelines is identified for the design of very high-density digital systems using inherently unreliable, error-prone devices.

Cotofana, *et al.* [17] introduced a design methodology that allows the system/circuit designer to build reliable systems out of unreliable nano-scaled components. The central point of the proposed approach is a generic (parametrical) architectural template, which is named configurable nanostructures for reliable Nano electronics (CONAN). CONAN embeds support for reliability at various levels of abstractions. Han & Jonker [22] developed the probability of the system survival of a defect-, and fault-tolerant architecture for nano-computers.

Motivated by the need for economical fault-tolerant designs for nano-architectures, Roy & Beiu [64] explored a novel multiplexing-based redundant design scheme with redundancy factors, R, at small ($R \leq 100$) and very small ($R \leq 10$) levels. In particular, the authors adapted a strategy known as von Neumann multiplexing to circuits of majority gates with three inputs, and analyzed the performance of a multiplexing scheme for very small redundancies using combinatorial arguments.

VI. Conclusions

The behavior of nano-scaled products is much more sensitive to changes in material compositions, manufacturing controllable variables, and noise parameters. This paper has reviewed various aspects of reliability research in the emerging nanotechnology. Around 80 papers from nearly 40 leading journal in nano-related areas have been covered, including 10 review papers summarized in Section IV.

We have broken down our presentation into the following four main topics:

- Introduction of concepts, and technical terms of reliability to nano-technology.
- Identification of physical failure mechanisms of nano-structured materials, and devices.
- Determination of quality parameters of nano-devices, failure modes, and failure analysis including reliability testing procedures, and instrumentation to localize nano-defects.
- Modeling of reliability functions, and failure rates of nano-systems.

Much work is needed in the nano-reliability field to ensure the product reliability, and safety in various use conditions.

References

[1] A. Agarwal, B. C. Paul, and K. Roy, "Process variation in nano-scale memories: Failure analysis and process tolerant architecture," presented at the Custom Integrated Circuits Conference 2004. Proceedings of the IEEE 2004, 2004, unpublished.

[2] S. J. Bae, S.-J. Kim, W. Kuo, and P. H. Kvam, "Statistical models for hot electron degradation in nano-scaled MOSFET devices," *IEEE Trans. Reliability, Special Issue on Reliability Studies on Nanotechnology*, 2007, In Press.

[3] A. H. Barber, I. Kaplan-Ashiri, S. R. Cohen, R. Tenne, and H. D. Wagner, "Stochastic strength of nanotubes: An appraisal of available data," *Composites Science and Technology*, vol. 65, pp. 2380–2384, 2005.

[4] C. Basaran, H. Tang, and S. Nie, "Experimental damage mechanics of microelectronic solder joints under fatigue loading," *Mechanics of Materials*, vol. 36, pp. 1111–1121, 2004.

[5] B. D. Beake and J. F. Smith, "Nano-impact testing-an effective tool for assessing the resistance of advanced wear-resistant coatings to fatigue failure and delamination," *Surface & Coatings Technology*, vol. 188–189, pp. 594–598, 2004.

[6] D. Bhaduri and S. Shukla, "NANOLAB—A tool for evaluating reliability of defect-tolerant nanoarchitectures," *IEEE Trans. Nanotechnology*, vol. 4, pp. 381–394, 2005.

[7] D. Bhaduri and S. K. Shukla, "NANOPRISM: A tool for evaluating granularity vs. reliability trade-offs in nano architectures," in *Proceedings of the 14th ACM Great Lakes Symposium on VLSI*, Boston, MA, USA, 2004.

[8] D. Bhaduri and S. K. Shukla, "Reliability analysis for defect-tolerant nano-architectures in the presence of interconnect noise," in *Nanotechnology, 2004. 4th IEEE Conference on*, 2004, pp. 602–604.

[9] H. J. L. Bressers, W. D. van Driel, K. M. B. Jansen, L. J. Ernst, and G. Q. Zhang, "From chemical building blocks of polymers to microelectronics reliability," presented at the Thermal and Mechanical Simulation and Experiments in Microelectronics and Microsystems, 2004. EuroSimE 2004. Proceedings of the 5th International Conference on, 2004, unpublished.

[10] A. Cester, S. Cimino, E. Miranda, A. Candelori, G. Ghidini, and A. Paccagnella, "Statistical model for radiation-induced wear-out of ultra-thin gate oxides after exposure to heavy ion irradiation," *IEEE Trans. Nuclear Science*, vol. 50, pp. 2167–2175, 2003.

[11] I. Chasiotis and A. McCarty, "Application of general Weibull statistics MEMS," *JOM*, vol. 56, pp. 193–194, 2004.

[12] S. Chatterjee, Y. Kuo, J. Lu, J. Y. Tewg, and P. Majhi, "Electrical reliability aspects of HfO2 high-k gate dielectrics with TaN metal gate electrodes under constant voltage stress," *Microelectronics Reliability*, vol. 46, pp. 69–76, 2006.

[13] M. X. Chen, X. J. Yi, Z. Y. Gan, and S. Liu, "Reliability of anodically bonded silicon-glass packages," *Sensors and Actuators A—Physical*, vol. 120, pp. 291–295, 2005.

[14] V. J. Chen and P. X. Ma, "The effect of surface area on the degradation rate of nano-fibrous poly(L-lactic acid) foams," *Biomaterials*, vol. 27, pp. 3708–3715, 2006.

[15] K. R. Cheung, "Ultrathin gate-oxide breakdown—Reversibility at low voltage," *IEEE Trans. Device and Materials Reliability*, vol. 6, pp. 67–74, 2006.

[16] S. H. Choa, "Reliability of MEMS packaging: Vacuum maintenance and packaging induced stress," *Microsystem Technologies-Micro-and Nanosystems-Information Storage and Processing Systems*, vol. 11, pp. 1187–1196, 2005.

[17] S. Cotofana, A. Schmid, Y. Leblebici, A. Ionescu, O. Soffke, P. Zipf, M. Glesner, and A. Rubio, "CONAN—A design exploration framework for reliable nano-electronics," presented at the Proceedings of the 2005 IEEE International Conference on Application-Specific Systems, Architecture Processors, 2005, unpublished.

[18] I. De Wolf, "MEMS reliability," *Microelectronics Reliability*, vol. 43, pp. 1047–1048, 2003.

[19] C. Finder, M. Beck, J. Seekamp, K. Pfeiffer, P. Carlberg, I. Maximov, F. Reuther, E. L. Sarwe, S. Zankovich, J. Ahopelto, L. Montelius, C. Mayer, and C. M. S. Torres, "Fluorescence microscopy for quality control in nanoimprint lithography," *Microelectronic Engineering*, vol. 67, pp. 623–628, 2003.

[20] R. Ganesan, T. K. Das, A. K. Sikder, and A. Kumar, "Wavelet-based identification of delamination defect in CMP (Cu-Low k) using nonstationary acoustic emission signal," *IEEE Trans. Semiconductor Manufacturing*, vol. 16, pp. 677–685, 2003.

[21] G. Groeseneken, R. Degraeve, B. Kaczer, and P. Roussel, "Recent trends in reliability assessment of advanced CMOS technologies," presented at the Microelectronic Test Structures, 2005. ICMTS 2005. Proceedings of the 2005 International Conference on, 2005, unpublished.

[22] J. Han and P. Jonker, "A defect- and fault-tolerant architecture for nanocomputers," *Nanotechnology*, vol. 14, pp. 224–230, 2003.

[23] K. Holmberg, "Reliability aspects of tribology," *Tribology International*, vol. 34, pp. 801–808, 2001.

[24] M.-H. Hsieh and S.-L. Jeng, "Accelerated discrete degradation models for leakage current of ultra-thin gate oxides," *IEEE Trans. Reliability, Special Issue on Reliability Studies on Nanotechnology*, 2007, In Press.

[25] M. Huff, "About MEMS and nanotechnology," [Online]. Available: http://www.memsnet.org/mems/

[26] ITRS, "International technology roadmap for semiconductors," [Online]. Available: http://www.itrs.net/reports.html

[27] Y. C. Jean, J. J. Zhang, H. M. Chen, Y. Li, and G. Liu, "Positron annihilation spectroscopy for surface and interface studies in nanoscale polymeric films," *Spectrochimica Acta Part A—Molecular and Biomolecular Spectroscopy*, vol. 61, pp. 1683–1691, 2005.

[28] J. Keller, A. Gollhardt, D. Vogel, and B. Michel, "Nanoscale deformation measurements for reliability analysis of sensors," presented at the Proceedings of the SPIE—The International Society for Optical Engineering, 2005, unpublished.

[29] M. D. Ker and K. H. Lin, "Overview on electrostatic discharge protection designs for mixed-voltage I/O interfaces: Design concept and circuit implementations," *IEEE Trans. Circuits and Systems I-Regular Papers*, vol. 53, pp. 235–246, 2006.

[30] K. O. Kim, W. Kuo, and W. Luo, "A relation model of gate oxide yield and reliability," *Microelectronics Reliability*, vol. 44, pp. 425–434, 2004.

[31] K. S. Kim, B. K. Cho, and T. W. Kim, "Switching and reliability issues of magnetic tunnel junctions for high-density memory device," *Current Applied Physics*, 2006, In Press.
[32] T. Kitamura, Y. Umeno, and A. Kushima, "Ideal strength of nano-components," *Materials Science Forum*, vol. 482, pp. 25–32, 2005.
[33] J.-H. Klein-Wiele, J. Bekesi, J. Ihlemann, and P. Simon, "Nanofabrication of solid materials with ultraviolet femtosecond pulses," *Proceedings of the SPIE—The International Society for Optical Engineering*, vol. 5399, pp. 139–146, 2004.
[34] A. M. Korsunsky and A. Constantinescui, "Work of indentation approach to the analysis of hardness and modulus of thin coatings," *Materials Science and Engineering a-Structural Materials Properties Microstructure and Processing*, vol. 423, pp. 28–35, 2006.
[35] P. Kouvelis and S. K. Mukhopadhyay, "Modeling the design quality competition for durable products," *IIE Transactions*, vol. 31, pp. 865–880, 1999.
[36] V. Kovacevic, S. Lucic, and M. Leskovac, "Morphology and failure in nanocomposites. Part I: Structural and mechanical properties," *Journal of Adhesion Science and Technology*, vol. 16, pp. 1343–1365, 2002.
[37] J. Krim, "Friction at the nano-scale," *Physics World*, vol. 18, pp. 31–34, 2005.
[38] H. Kufluoglu and M. A. Alam, "Theory of interface-trap-induced NBTI degradation for reduced cross section MOSFETs," *IEEE Trans. Electron Devices*, vol. 53, pp. 1120–1130, 2006.
[39] C. J. M. Lasance, "Thermally driven reliability issues in microelectronic systems: Status-quo and challenges," *Microelectronics Reliability*, vol. 43, pp. 1969–1974, 2003.
[40] Y. C. Lee, B. Amir Parviz, J. A. Chiou, and C. Shaochen, "Packaging for microelectromechanical and nanoelectromechanical systems," *Advanced Packaging, IEEE Trans. [see also Components, Packaging and Manufacturing Technology, Part B: Advanced Packaging, IEEE Trans.]*, vol. 26, pp. 217–226, 2003.
[41] Y. M. Lee, Y. D. Wu, and G. Lucovsky, "Breakdown and reliability of p-MOS devices with stacked RPECVD oxide/nitride gate dielectric under constant voltage stress," *Microelectronics Reliability*, vol. 44, pp. 207–212, 2004.
[42] Y. Li and C. P. Wong, "Recent advances of conductive adhesives as a lead-free alternative in electronic packaging: Materials, processing, reliability and applications," *Materials Science & Engineering R—Reports*, vol. 51, pp. 1–35, 2006.
[43] W. C. Lin, T. C. Wu, Y. H. Tsai, L. J. Du, and Y. C. King, "Reliability evaluation of class-E and class-A power amplifiers with nanoscaled CMOS technology," *IEEE Trans. Electron Devices*, vol. 52, pp. 1478–1483, 2005.
[44] Y. Liu, A. Sadat, and J. S. Yuan, "Gate oxide breakdown on nMOSFET cutoff frequency and breakdown resistance," *IEEE Trans. Device and Materials Reliability*, vol. 5, pp. 282–288, 2005.
[45] S. Lombardo, J. H. Stathis, B. P. Linder, K. L. Pey, F. Palumbo, and C. H. Tung, "Dielectric breakdown mechanisms in gate oxides," *Journal of Applied Physics*, vol. 98, 2005, pp..
[46] W. Luo, D. Sunardi, Y. Kuo, and W. Kuo, "Stress testing and characterization of high-k dielectric thin films," presented at the Integrated Reliability Workshop Final Report, 2003 IEEE International, 2003, unpublished.
[47] W. Luo, T. Yuan, Y. Kuo, J. Lu, J. Yan, and W. Kuo, "Breakdown phenomena of zirconium-doped hafnium oxide high-k stack with an inserted interface layer," *Applied Physics Letters*, vol. 89, 2006, pp..
[48] W. Luo, T. Yuan, Y. Kuo, J. Lu, J. Yan, and W. Kuo, "Charge trapping and dielectric relaxation in connection with breakdown of high-k gate dielectric stacks," *Applied Physics Letters*, vol. 88, 2006, pp..
[49] P. H. Mayrhofer, C. Mitterer, L. Hultman, and H. Clemens, "Microstructural design of hard coatings," *Progress in Materials Science*, vol. 51, pp. 1032–1114, 2006.
[50] S. Melle, D. De Conto, D. Dubuc, K. Grenier, O. Vendier, J. L. Muraro, J. L. Cazaux, and R. Plana, "Reliability modeling of capacitive RF MEMS," *IEEE Trans. Microwave Theory and Techniques*, vol. 53, pp. 3482–3488, 2005.
[51] B. Michel, "Micro reliability and lifetime estimation," [Online]. Available: http://www.pb.izm.fhg.de/izm/015_Programms/05_Micro/index.html
[52] E. Miranda and D. Jimenez, "A new model for the breakdown dynamics of ultra-thin gate oxides based on the stochastic logistic differential equation," presented at the Proceedings of the 24th International Conference on Microelectronics 2004, unpublished.
[53] P. H. F. Morshuis, "Degradation of solid dielectrics due to internal partial discharge: Some thoughts on progress made and where to go now," *IEEE Trans. Dielectrics and Electrical Insulation*, vol. 12, pp. 905–913, 2005.
[54] T. Namazu and Y. Isono, "High-cycle fatigue test of nanoscale Si and SiO2 wires based on AFM technique," presented at the Micro Electro Mechanical Systems, 2003. MEMS-03 Kyoto. IEEE The Sixteenth Annual International Conference on, 2003, unpublished.
[55] G. Norman, D. Parker, M. Kwiatkowska, and S. K. Shukla, "Evaluating the reliability of defect-tolerant architectures for nanotechnology with probabilistic model checking," presented at the Proceedings of the 17th International Conference on VLSI Design, 2004, unpublished.
[56] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, Calif: Morgan Kaufmann Publishers, 1988.
[57] S. Peng and K. J. Cho, "Ab initio study of doped carbon nanotube sensors," *Nano Letters*, vol. 3, pp. 513–517, 2003.
[58] F. Pervin, Y. X. Zhou, V. K. Rangari, and S. Jeelani, "Testing and evaluation on the thermal and mechanical properties of carbon nano fiber reinforced SC-15 epoxy," *Materials Science and Engineering A—Structural Materials Properties Microstructure and Processing*, vol. 405, pp. 246–253, 2005.
[59] A. Prabhakumar, J. Campbell, R. Mills, P. Gillespie, D. Esler, S. Rubinsztajn, S. Tonapi, and K. Srihari, "Assemble and reliability of flip chips with a nano-filled wafer level underfill," presented at the Electronics Packaging Technology Conference, 2004, unpublished.
[60] N. Pugno, B. Peng, and H. D. Espinosa, "Predictions of strength in MEMS components with defects—A novel experimental-theoretical approach," *International Journal of Solids and Structures*, vol. 42, pp. 647–661, 2005.
[61] G. Ribes, J. Mitard, M. Denais, S. Bruyere, F. Monsieur, C. Parthasarathy, E. Vincent, and G. Ghibaudo, "Review on high-k dielectrics reliability issues," *IEEE Trans. Device and Materials Reliability*, vol. 5, pp. 5–19, 2005.
[62] R. O. Ritchie, C. L. Muhlstein, and R. K. Nalla, "Failure by fracture and fatigue in "nano" and ''bio" materials," *International Journal Series A—Solid Mechanics and Material Engineering*, vol. 47, pp. 238–251, 2004.
[63] M. C. Roco, "International strategy for nanotechnology research and development," *Journal of Nanoparticle Research*, vol. 3, pp. 353–360, 2001.
[64] S. Roy and V. Beiu, "Majority multiplexing-economical redundant fault-tolerant designs for nanoarchitectures," *IEEE Trans. Nanotechnology*, vol. 4, pp. 441–451, 2005.
[65] A. Schmid and Y. Leblebici, "Robust circuit and system design methodologies for nanometer-scale devices and single-electron transistors," *IEEE Trans. Very Large Scale Integration (VLSI) Systems*, vol. 12, pp. 1156–1166, 2004.
[66] M. Schrenk, K. Koller, K. H. Allers, and H. Korner, "Integration of metal insulator metal capacitors (MIM-Caps) with low defectivity into a copper metallization," *Microelectronic Engineering*, vol. 82, pp. 514–520, 2005.
[67] U. Schwalke, M. Polzl, T. Sekinger, and M. Kerber, "Ultra-thick gate oxides: Charge generation and its impact on reliability," *Microelectronics Reliability*, vol. 41, pp. 1007–1010, 2001.
[68] A. Sikora, F. P. Pesl, W. Unger, and U. Paschen, "Technologies and reliability of modern embedded flash cells," *Microelectronics Reliability*, vol. 46, pp. 1980–2005, 2006.
[69] G. Sivalingam and G. Madras, "Photocatalytic degradation of poly(bisphenol-A-carbonate) in solution over combustion-synthesized TiO2: Mechanism and kinetics," *Applied Catalysis A—General*, vol. 269, pp. 81–90, 2004.
[70] V. T. Srikar and S. D. Senturia, "The reliability of microelectromechanical systems (MEMS) in shock environments," *Journal of Microelectromechanical Systems*, vol. 11, pp. 206–214, 2002.
[71] J. H. Stathis, "Reliability limits for the gate insulator in CMOS technology," *IBM Journal of Research and Development*, vol. 46, pp. 265–286, 2002.
[72] F. Su, Y. Sun, X. Shi, and S. Wong Chee Khuen, "Techniques for nanoscale deformation measurement," presented at the Electronics Packaging Technology Conference, 2004. EPTC 2004. Proceedings of 6th, 2004, unpublished.
[73] J. S. Suehle, B. Zhu, Y. Che, and J. B. Bernstein, "Acceleration factors and mechanistic study of progressive breakdown in small area ultra-thin gate oxides," presented at the Reliability Physics Symposium Proceedings, 2004. 42nd Annual. 2004 IEEE International, 2004, unpublished.
[74] N. Tomczak, R. A. L. Vallee, E. M. H. P. van Dijk, M. Garcia-Parajo, L. Kuipers, N. F. van Hulst, and G. J. Vancso, "Probing polymers with single fluorescent molecules," *European Polymer Journal*, vol. 40, pp. 1001–1011, 2004.

[75] T. Umemura, Y. Mizutani, T. Okamoto, T. Taguchi, K. Nakajima, and K. Tanaka, "Life expectancy and degradation behavior of electric double layer capacitor part I," presented at the Properties and Applications of Dielectric Materials, 2003. Proceedings of the 7th International Conference on, 2003, unpublished.
[76] D. P. Vallett, "Failure analysis requirements for nanoelectronics," *IEEE Trans. Nanotechnology*, vol. 1, pp. 117–121, 2002.
[77] W. M. van Spengen, "MEMS reliability from a failure mechanisms perspective," *Microelectronics Reliability*, vol. 43, pp. 1049–1060, 2003.
[78] H. Wong and H. Iwai, "On the scaling issues and high-kappa replacement of ultrathin gate dielectrics for nanoscale MOS transistors," *Microelectronic Engineering*, vol. 83, pp. 1867–1904, 2006.
[79] E. B. Zeynalov and N. S. Allen, "Effect of micron and nano-grade titanium dioxides on the efficiency of hindered piperidine stabilizers in a model oxidative reaction," *Polymer Degradation and Stability*, vol. 91, pp. 931–939, 2006.
[80] C. Zhao, X. Bai, and S. Dey, "Evaluating transient error effects in digital nanometer circuits," *IEEE Trans. Reliability, Special Issue on Reliability Studies on Nanotechnology*, 2007, In Press.
[81] L. Zhihong, Z. Weiquan, and M. Fuchen, "Build-in reliability analysis for circuit design in the nanometer technology era," presented at the Integrated Circuit Design and Technology, 2004. ICICDT '04. International Conference on, 2004, unpublished.

Shuen-Lin Jeng is an Associate Professor in the Department of Statistics at National Cheng Kung University, Taiwan. He received his M.S. degree in Applied Mathematics from Fujen University, Taiwan, in 1989; and Ph.D. in Statistics from Iowa State University, U.S.A, in 1998. He is an elected member of the International Statistical Association (ISA), a member of American Statistical Association (ASA), Chinese Institute of Probability and Statistics (CIPS), and Chung-hua Data Mining Society (CDMS). His research interests include product reliability, software reliability, statistical computing, industrial statistics, and data mining.

Jye-Chyi Lu received his Ph.D. degree in statistics in 1988 at the University of Wisconsin, and was a professor in the Department of Statistics at North Carolina State University from 1988 to 1999. Now, he is a professor in the School of Industrial and Systems Engineering at Georgia Institute of Technology. Dr. Lu has about 55 disciplinary, and interdisciplinary publications appearing in both engineering, and statistics journals. Currently, he is an Associate Editor for *Technometrics*, IEEE TRANSACTIONS ON RELIABILITY, and *Journal of Quality Technology.* His research areas cover industrial statistics, signal processing, semiconductor and electronic manufacturing, data mining, and a few new topics such as bioinformatics, supply-chain management, logistics planning, and nanotechnology.

Kaibo Wang is an Assistant Professor in the Department of Industrial Engineering, Tsinghua University, Beijing, P. R. China. He received his B.S., and M.S. degrees in Mechatronics from Xi'an Jiaotong University, Xi'an, P.R. China; and his Ph.D. in Industrial Engineering and Engineering Management from the Hong Kong University of Science and Technology, Hong Kong. He is an ASQ Certified Six Sigma Black Belt. His research interests include quality engineering and management, statistical process control, and statistical methods for nanotechnology.

Molecular dynamics simulation of nanomaterials using an artificial neural net

Mark Benedict and John F. Maguire*
Air Force Research Laboratory, Polymers Branch/ MLBP, Dayton, Ohio 45433-7746, USA
(Received 22 July 2003; revised manuscript received 4 February 2004; published 23 November 2004)

We report a method of conducting molecular dynamics (MD) simulations that uses an artificial neural net (ANN) to significantly increase computational speed. The technique enables dynamical simulation of hard objects with essentially arbitrarily complex geometry and is well suited to the simulation of granular matter over a wide range of densities. In hard systems, binary collisions are well defined and the ANN approach enables an efficient algorithm to determine the time to next collision with high accuracy. The method has been used to enable an MD study of an ensemble of 1800 hard, smooth, impenetrable equilateral triangles in a two-dimensional periodic space. At high packing fraction ($0.6<\rho<0.9$), the hard-triangle system exists as a liquid-crystalline-like phase (LCP) in which there is no long-range translational order but in which there is nearly perfect long-range orientational order. As the packing fraction decreases, the LCP undergoes a transition to a fluid state in which the long-range orientational correlation vanishes but short-range order is retained. Long-lived clusters, notably hexamers, are clearly apparent in the liquid phase and appear to be stabilized by a sort of internal "orientational" osmotic pressure. Insofar as can be inferred from our machine calculations, the transition between the LCP and the liquid occurs around $\rho\sim0.57$ and appears to be second order. At low density, the hard-triangle system undergoes "chattering" collisions in which pairs of triangles collide and become associated, undergoing multiple collisions with each other before colliding with a third particle. The radial distribution function obtained from both molecular dynamics and Monte Carlo calculations shows a weak peak at low packing fraction.

DOI: 10.1103/PhysRevB.70.174112 PACS number(s): 61.20.Ja, 61.30.Gd

I. INTRODUCTION

Computer simulation has played an important role in developing a better understanding of the physics of dense, disordered media.[1,2] The study of particles that interact only through infinitely repulsive forces at the point of contact (hard systems) and with no frictional losses is of special interest in that the interparticle potential energy is defined exactly and there are no three-body or higher order terms. Modern theories of the liquid state are based on the idea that a hard-sphere fluid can serve as a reference system from which the properties of more realistic liquids can be obtained by perturbation theory.[3–5] Alder and Wainright[6] published the first paper on the molecular dynamics of the hard-sphere system almost 50 years ago and studies of the *N*-body properties of hard spheres continue to yield important insights[7] into the fundamental mechanisms of phase transitions, nucleation, and crystal growth.

Of course, not all objects are spherical and there has been considerable recent interest in the behavior of systems in which the interparticle potential is nonspherical. For example, in a series of elegant experiments, Whitesides *et al.*[8,9] have demonstrated clearly the importance of geometrical considerations in the development of long-range structure and have suggested how such structures might be used to engineer nanomaterials. Recently, de Wild *et al.*[10] have used molecules with triangular symmetry (subphthalocyanines) to produce a number of long-range structures on gold surfaces. Not surprisingly, mixtures of triangles and spheres give rise to rich phase behavior, with triangles packing locally to form hexagons in some phases and linear chains in other phases. In all of this work an important question arises; namely, to what degree are the observed phases a consequence of the purely topological and space-filling attributes of the particle and to what degree might they be a function of chemical interaction?

In order to answer questions of this sort we need to have an ability to model large ensembles of hard space-filling or essentially granular objects of arbitrarily complex geometry. Granular materials are intrinsically interesting and it would seem plausible that the shape of grains of sand or of snowflakes may be an important factor in phenomena ranging from the structure of planetary rings to avalanche. Furthermore, if the topological constraints imposed by geometry can be taken into account explicitly and the properties of the reference system obtained by simulation, it may be possible to extend analytic perturbation methods to describe real materials such as complex fluids, polymers, polycrystalline metals, and heaps of sand or snow, etc. In this way, the simulation methods described here, initially for hard granular materials, may provide a pathway to circumvent the serious limitations of time scale and length scale that limit the usefulness of current molecular dynamics simulations in complex systems.

On the experimental side, it is exceedingly difficult to prepare and reproduce, at different laboratories, granular materials in a well defined state in which the interparticle interactions (coefficient of restitution, shape distribution, boundary conditions, water content, etc.) are adequately specified. Rather large error bars are common even within a given laboratory. It is in precisely such circumstances that an ability to obtain exact, albeit numerical, results on realistic (i.e., prickly) well defined granular systems in an identifiable ensemble may prove especially useful.

Notwithstanding the above comments, there has been comparatively little work on the dynamics of hard space-

filling bodies of nonspherical geometry. Frenkel and Maguire[11] and Magda *et al.*[12] have reported the transport properties of a fluid of infinitely thin hard-line segments. More recently Aspelmeier *et al.* have used their approach to simulate granular cooling of hard needles,[13] while Yoshimura and Mukoyama[14] have made detailed studies of isolated binary chattering collisions between rods. While infinitely thin rods exhibit many interesting transport properties, the thermodynamics is uninteresting since they do not fill space. These calculations are exceedingly intensive in terms of computer time, and it is perhaps for this reason that the studies have not heretofore been extended to hard space-filling platonic solids.

In this paper we report a practical method for molecular dynamics simulation of large ensembles of space-filling hard nonspherical objects. The algorithm is based on an application of artificial neural nets (ANNs) and runs at execution speeds that are of the same order as those of hard spheres. The approach is illustrated for the simplest example: the smooth, impenetrable, hard triangle of uniform mass distribution in two dimensions. Interestingly, even this simplest example reveals intriguing phenomena.

II. COLLISION DETECTION

Details of the molecular dynamics (MD) method have been presented in a number of excellent recent books,[1,2] and only those aspects of relevance in the extension of the technique to the simulation of platonic solids will be discussed here. In order to conduct an MD simulation of hard objects it is necessary to predict when and where two particles in the ensemble will next collide. Given the discontinuous interaction potential of the rigid body model, no forces act on the particles until they come into contact, so the minimum of the list of collision times defines the time to next collision. For spheres this requires the solution of a quadratic equation, which can be accomplished very rapidly with a digital computer. For nonspherical objects the situation is more involved and requires a time consuming iterative search for a particular root of an oscillatory transcendental function.[11] The idea behind the present work is very simple. Rather than solve a transcendental equation in the inner loop of an MD calculation, it is far more efficient to train an ANN[15] over the Hilbert space of positions and momenta available to each pair of potential collision partners. The Hilbert space can be populated to any required degree of accuracy and matters can be arranged so that extrapolation outside the space of exemplars is never required. This procedure can be applied to objects of essentially arbitrarily complex shape, but is illustrated here for the simple example of hard, impenetrable equilateral triangles in two dimensions. Each triangle has unit mass and uniform mass distribution.

Using the notation shown in Fig. 1, the position of triangles i and j may be written

$$\vec{r}_i(t) = \vec{r}_i(0) + \vec{v}_i(t)\Delta t, \tag{1}$$

$$\vec{r}_j(t) = \vec{r}_j(0) + \vec{v}_j(t)\Delta t, \tag{2}$$

where t is time and $\vec{v}$ and $\vec{r}$ are the velocities and positions respectively. The orientation is defined by the unit vector $\hat{u}(t)$, as

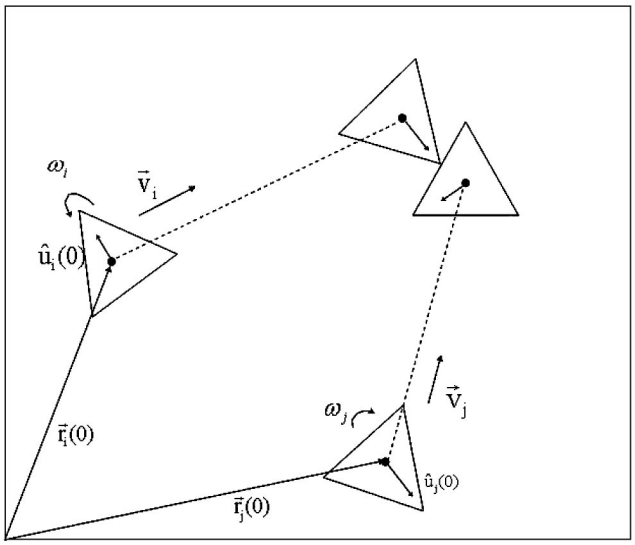


FIG. 1. Schematic illustration of the collision of a pair of triangles. Positions are represented by $\vec{r}_i$ and $\vec{r}_j$, velocities by $\vec{V}_i$ and $\vec{V}_j$, orientations by $\hat{u}_i$ and $\hat{u}_j$, and angular velocities by ω_i and ω_j.

$$\hat{u}(t) = \hat{u}(0)\cos(\omega t) + \frac{\vec{J}}{|J|} \times \hat{u}(0)\sin(\omega t), \tag{3}$$

where ω is the angular velocity and $\vec{J}$ is the angular momentum.

The condition for a collision may be written

$$f(t) = \left\{ \left[r_i(t_c) + R\left(n_1 \frac{2\pi}{3}\right) u_i(t_c) \right] - \left[r_j(t_c) + R\left(n_3 \frac{2\pi}{3}\right) u_j(t_c) \right] \right\} \times \left\{ \left[r_i(t_c) + R\left(n_2 \frac{2\pi}{3}\right) u_i(t_c) \right] - \left[r_j(t_c) + R\left(n_3 \frac{2\pi}{3}\right) u_j(t_c) \right] \right\}, \tag{4}$$

subject to the conditions that when $f(t_c)=0$, the point of contact lies on the line segment between the vertices, and $f(t_c)$ is computed iteratively for all permutations of $n_1 \neq n_2$ on the interval $[0,2]$ where R denotes the rotation operator. This represents the combination of all sides and vertices for a given pair of triangles.

Equation (4) may be solved numerically for the time to next collision t_c using a straightforward modification of the *a posteriori* collision detection method first described by Robertus and Sando.[16] In the usual implementation of the *posteriori* detection, the overlap function is determined as a function of time using a constant time interval until a condition of overlap is detected. The particle positions are then taken back one time step, the time step is reduced, usually halved, and the procedure repeated until the time to collision is computed to the desired accuracy. The difficulty with the approach is that for any finite time step it is in principle possible to encounter a pair of particles that will pass undetected through each other. If this happens, the overlap condition that is eventually detected will be for the wrong pair of particles

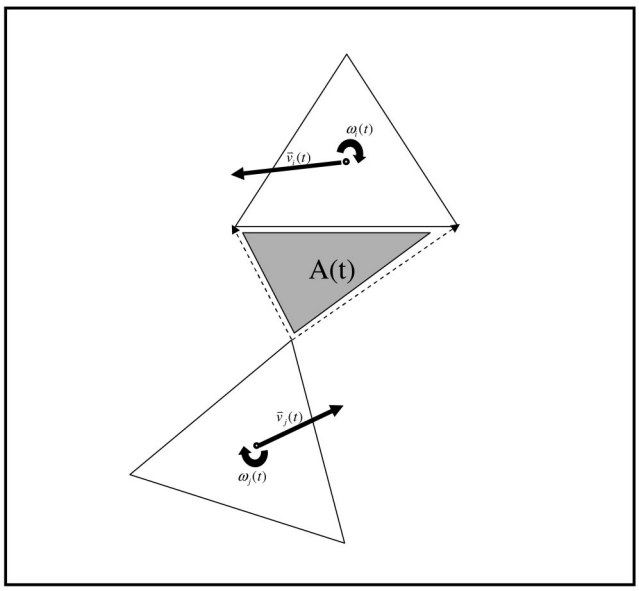


FIG. 2. Schematic representation of the method used to generate the exemplars for neural net training. The collision time is defined as the time at which the area $A(t)$ is equal to zero subject to the auxiliary condition that the point of contact lies within the line segment that defines the triangle. $A(t)$ is calculated for all combinations of sides and vertices.

which then exchange momenta, while conserving energy, and all will appear well with a simulation that is seriously flawed. This can be a severe problem particularly with particles that have sharp points. It is possible to avoid this state of affairs by considering the relative linear and angular velocities and by conducting the simulation using a range of time steps and tests to assure that this type of glancing transient penetration is avoided. Within the inner loop of an MD simulation such an approach is very time consuming and renders the approach impractical. However, this kind of problem is well suited to a solution using an ANN.[17] At a given time, the positions, velocities, and angular velocities of a pair of triangles in two dimensions may be represented by a vector in seven-dimensional Hilbert space. This dimensionality is only moderately large[15] and is easily manageable with current ANN algorithms. For a single pair of triangles, training data were generated over 10^6 configurations spanning essentially the complete range (three standard deviations) in both the translational and rotational distributions. This was accomplished as illustrated in Fig. 2 using an iterative numeric solution for each pair of side of the triangles. The collision time is taken as the time at which $f(t) < 10^{-12}$. In the present work we have used the easily differentiable log-sigmoid as the transfer function in a back-propagation net.[17] While these log-sigmoid functions are certainly not optimized for the present problem, it will be appreciated that training time is relatively insignificant and no great effort need be expended in this regard.

Training the net using a back-propagation algorithm and 10^6 exemplars requires about 48 h of CPU time on an eight-node (2 GHz) cluster. Once the net is trained, t_c is returned with about the speed of a hard-sphere calculation, all other

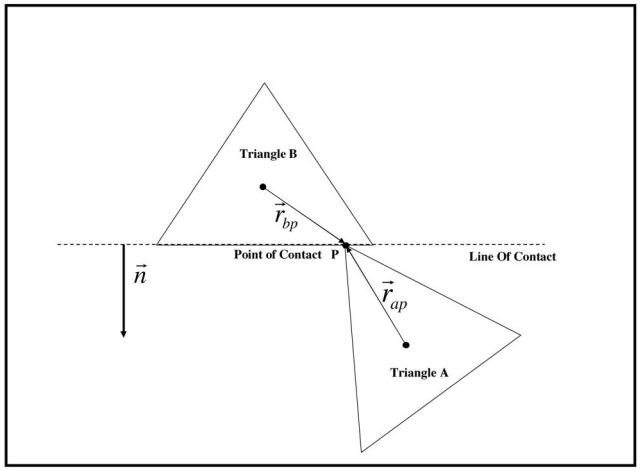


FIG. 3. Illustration of the collision dynamics for a pair of triangles showing the lines of contact and the direction of momentum transfer.

parameters being equal. This is a significant advantage in that it theoretically enables simulations of objects of arbitrarily complex geometry with speeds that are comparable to those for hard spheres. It will, of course, be appreciated that particles of highly complex geometries require testing of all combinations of sides, edges, and vertices, and this requires a combinatorial increase[18] in training time for the ANN. Moreover, while the net has sufficiently high precision over most of the density range, the direct *a posteriori* method must be used to obtain the highest precision in the event that the net returns t_c values for multiple pairs that are exceedingly close. This is not a significant problem at low and moderate density, but becomes more troublesome as the density approaches close packed. Use of the neural net in this context amounts essentially to applying sophisticated fitting and interpolation procedures in order to fit a comparatively

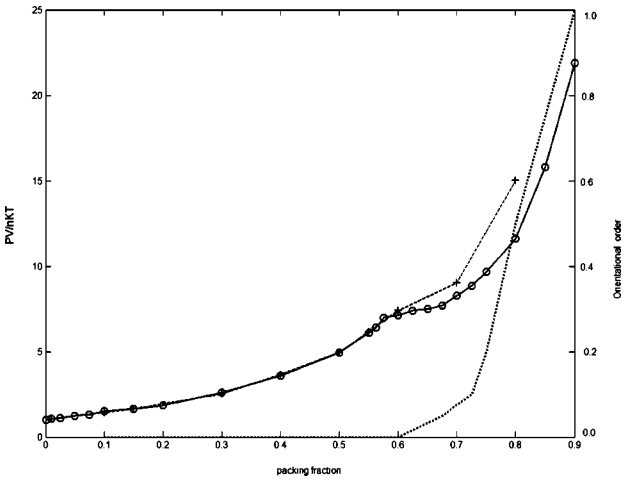


FIG. 4. Equation of state for hard triangles in two dimensions. The open circles connected by the solid line correspond to runs in an expansion cycle, while the crosses connected by the dashed line correspond to runs in a compression cycle. The dotted line shows the hexagonal orientational order factor g_6 as a function of packing fraction.

sparse but very precise data set in a high-dimensional space. Another way to view the use of the neural net in this application is to view it as an efficient filter that reduces reliance upon *a posteriori* methods so that large amounts of computer time can be saved by filtering out pairs that do not have times to collision that are within a narrow window of the global minimum. In this sense the ANN is of little or no scientific interest *per se*, but it does allow the algorithm to perform at up to an order of magnitude better than the *a posteriori* method speed, and in turn enables the exploration of complex phenomena that depend on the collective dynamics of the many-body system.

III. DYNAMICS

The velocities at the point of contact may be written

$$\vec{v}_{ip} = \vec{v}_i + \omega_i \times \vec{r}_{ip}, \tag{5}$$

$$\vec{v}_{jp} = \vec{v}_j + \omega_j \times \vec{r}_{jp}, \tag{6}$$

where $\vec{v}_{ip}$ is the velocity of point P on triangle i prior to collision, $\vec{v}_i$ is the translational (center of mass) velocity of triangle i prior to collision, ω_i is the rotational velocity of triangle i about its center prior to collision, and $\vec{r}_{ip}$ is the vector from the center of mass of triangle i to the point of contact P.

The assumption that the triangles are perfectly smooth is equivalent to neglecting the components of velocity that are parallel to the line of contact; hence, the relative velocity at the point of contact can be expressed as

$$\vec{v}_{rel} = \vec{n} \cdot (\vec{v}_{ip} - \vec{v}_{jp}), \tag{7}$$

$$\vec{v}'_{rel} = \vec{n} \cdot (\vec{v}'_{ip} + \vec{v}'_{jp}), \tag{8}$$

with $\vec{n}$ being the normal to the line of contact in the direction of momentum transfer. The impulse is

$$\vec{J} = \Delta(m\vec{v}), \tag{9}$$

and because the triangles are assumed to be smooth and hard, only momentum components normal to the line of contact will be transferred, as

$$\vec{J} = J \cdot \vec{n}. \tag{10}$$

The velocity changes on a rigid body of mass m and moment of inertia I due to this impulse are given by

$$\Delta\vec{v} = \frac{j \cdot \vec{n}}{m} \quad \Delta\omega = \frac{(\vec{r} \times j\vec{n})}{I}. \tag{11}$$

For triangle i the postcollision velocity is

$$\vec{v}'_i = \vec{v}_i + \frac{j\vec{n}}{m_i}, \quad \omega'_i = \omega_i + \frac{(\vec{r}_{ip} \times \vec{n})}{I_i}. \tag{12}$$

Combining these terms and calculating the linear velocity at the point of contact,

$$\vec{v}'_{ip} = \left(\vec{v}_i + \frac{j\vec{n}}{m_i}\right) + \left(\omega_i + \frac{\vec{r}_{ip} \times \vec{n}}{I_i}\right) \times \vec{r}_{ip}, \tag{13}$$

or in a form that is a linear function of the impulse j, as

$$\vec{v}'_{ip} = \vec{v}_{ip} + j\left(\frac{n}{m_i} + \frac{\vec{r}_{ip} \times \vec{n}}{I_i}\right) \times \vec{r}_{ip}. \tag{14}$$

The opposite impulse acts on triangle j, as

$$\vec{v}'_{jp} = \vec{v}_{jp} - j\left(\frac{n}{m_j} + \frac{\vec{r}_{jp} \times \vec{n}}{I_j}\right) \times \vec{r}_{jp}, \tag{15}$$

so that the postcollision relative velocity at the point of contact can be expressed as

$$\vec{v}'_{rel} = \vec{n} \cdot (\vec{v}'_{ip} - \vec{v}'_{jp}), \tag{16}$$

$$\vec{v}'_{rel} = \vec{n} \cdot (\vec{v}_{ip} - \vec{v}_{jp}) + j\left[\frac{\vec{n} \cdot \vec{n}}{m_i} + \frac{\vec{n} \cdot \vec{n}}{m_j} + \vec{n} \cdot \left(\frac{\vec{r}_{ip} \times \vec{n}}{I_i}\right) \times \vec{r}_{ip} + \vec{n} \cdot \left(\frac{\vec{r}_{jp} \times \vec{n}}{I_j}\right) \times \vec{r}_{jp}\right], \tag{17}$$

yielding

$$\vec{v}'_{rel} = \vec{v}_{rel} + j\left[\frac{1}{m_i} + \frac{1}{m_j} + \vec{n} \cdot \left(\frac{\vec{r}_{ip} \times \vec{n}}{I_i}\right) \times \vec{r}_{ip} + \vec{n} \cdot \left(\frac{\vec{r}_{jp} \times \vec{n}}{I_j}\right) \times \vec{r}_{jp}\right]. \tag{18}$$

Now that the postcollision relative momentum for these two bodies has been written in terms of the linear momentum transfer at the point of contact, application of the conservation of linear momentum yields

$$m_i\vec{v}'_{ip} + m_j\vec{v}_{jp} = m_i\vec{v}_{ip} + m_j\vec{v}'_{jp}, \tag{19}$$

$$\vec{v}'_{rel} = -\vec{v}_{rel}. \tag{20}$$

Substituting back into the above equation gives

$$-\vec{v}'_{rel} = \vec{v}_{rel} + j\left[\frac{1}{m_i} + \frac{1}{m_j} + \vec{n} \cdot \left(\frac{\vec{r}_{ip} \times \vec{n}}{I_i}\right) \times \vec{r}_{ip} + \vec{n} \cdot \left(\frac{\vec{r}_{jp} \times \vec{n}}{I_j}\right) \times \vec{r}_{jp}\right]. \tag{21}$$

Solving for the impulse j and setting the coefficient of restitution to unity yields an expression[19] for the momentum that is transferred when the two triangles collide, given by

$$j = \frac{-2v_{rel}}{\left[\frac{1}{m_i} + \frac{1}{m_j} + \vec{n} \cdot \left(\frac{\vec{r}_{ip} \times \vec{n}}{I_i}\right) \times \vec{r}_{ip} + \vec{n} \cdot \left(\frac{\vec{r}_{jp} \times \vec{n}}{I_j}\right) \times \vec{r}_{jp}\right]}. \tag{22}$$

When the momentum transfer is known the pressure can be determined from the virial theorem or less straightforwardly by extrapolation of the angle-averaged radial distribution function at the point of contact[1] and integrating over the square of the separation at contact.

In discussing orientational order in these systems it is useful to define a parameter g_6 by the equation:

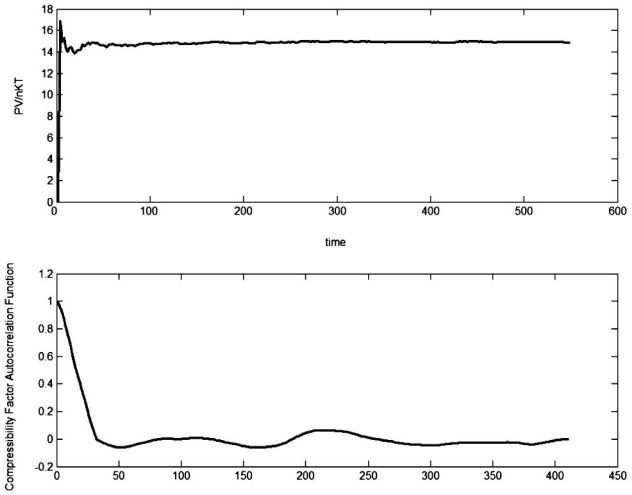


FIG. 5. (top) A plot of the compressibility factor as a function of time after the system has been expanded from $\rho=0.9$ to $\rho=0.85$. (bottom) The compressibility factor autocorrelation function as a function of time.

$$g_6 = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} e^{\frac{i\pi\theta_{ij}}{3}}, \tag{23}$$

where θ_{ij} is the angle between $\hat{u}_i$ and $\hat{u}_j$. For perfectly hexagonally ordered systems in two dimensions $g_6=1$, while for an orientationally random system $g_6=0$.

IV. RESULTS AND DISCUSSION

The equation of state is shown in Fig. 4. In this figure the open circles were obtained from runs that were initiated from a nearly close-packed solid and the system was undergoing expansion. The crosses were obtained when the system was undergoing compression from an initial very low density phase in which the positions and orientations of each triangle were chosen at random. In both cases the initial velocities were chosen from a Maxwellian distribution centered at zero and in the range of $[-1,1]$. On expansion the compressibility factor, $Z=PV/nKT$, drops smoothly until $\rho\sim0.57$, at which point a discontinuity in slope, indicative of a second-order transition, is observed. In these expansion runs the first 300 collisions per particle were discarded and the averages were taken typically over 700 to 1000 collisions per particle in order to assure, insofar as it is possible in a machine calculation, that the system had attained equilibrium. In the low density regime ($\rho<0.5$), the compression and expansion results lie on the same curve. On compression above $\rho\sim0.5$, there is evidence of hysteresis with overshoot into a glassy region and exceedingly slow relaxation.

We have paid particular attention to the question of thermodynamic equilibrium in these simulations. Figure 5(top) shows a plot of the compressibility factor as a function of time after the system had been expanded from $\rho=0.9$ to $\rho=0.85$. Note that the abscissa in this plot refers to the average number of collisions per particle for a system of 1800 particles; i.e., the run lasted approximately 1.5×10^6 collisions.

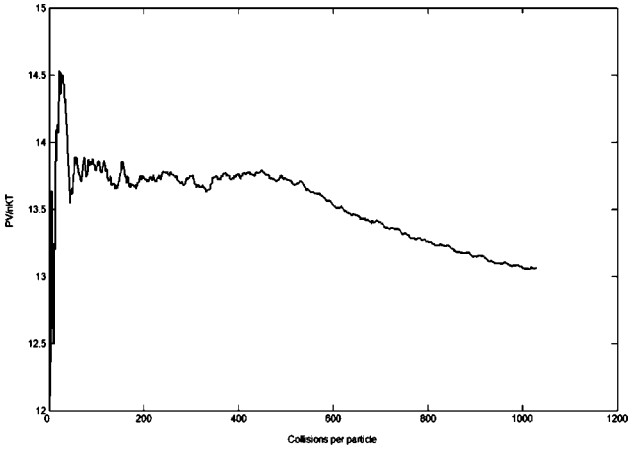


FIG. 6. Compressibility factor as a function of number of collisions per particle for a system at $\rho=0.80$. There is a very slow relaxation in the glassy region on compression. This slow feature is absent in the expansion cycle.

Clearly there is initial slow, collective, transient behavior that requires a long time ($\sim$50 collisions per particle) to relax. This portion was discarded when taking averages. Figure 5(bottom) shows the autocorrelation function for the fluctuation in compressibility factor. Here it is clearly apparent that the "production run" part of the simulations, i.e., the section of the simulations that are considered equilibrated and used to generate statistics, have been performed for at least eight times longer than the correlation time for the compressibility factor fluctuation. Moreover, as seen in Fig. 4, for the packing fractions in the interval $[0.1,0.5]$ identical results are obtained regardless of whether the system is prepared by expansion or compression; hence, it would appear by any usual measure that the solid line in Fig. 4 represents the equilibrium condition for this system. However, when the system is equilibrated at low packing fraction ($\rho<0.55$) and compressed through the transition point, it goes into a glassy region as represented by the dashed segment joining the crosses. In this regime the relaxation times become exceedingly long, as shown in Fig. 6. Here the relaxation time of the fluctuation of the compressibility factor is over 5000 collisions per particle, suggesting that a simulation of over 10×10^6 total collisions would be required to attain equilibrium following compression into the glassy region. Since we already have the equilibrium data from the expansion runs, it was thought that this aspect might be pursued in future studies.

Turning now to the question of the chattering collisions and the implications for equilibrium, the compressibility factor calculations have been carried out for very long times and the usual criteria for equilibrium have been met. When the packing fraction is $\sim\rho=0.7$, each triangle has just enough room to rotate freely on average. It is seen that, below $\rho=0.55$, each triangle has enough area on average to rotate and translate freely; i.e., we are in the liquid rather than the liquid crystal regime. Because of the triangular symmetry it is likely that when a pair of triangles collide, initially chatter for a time, and then move outside their area of mutual influence (the outscribed circle of each triangle), they are still

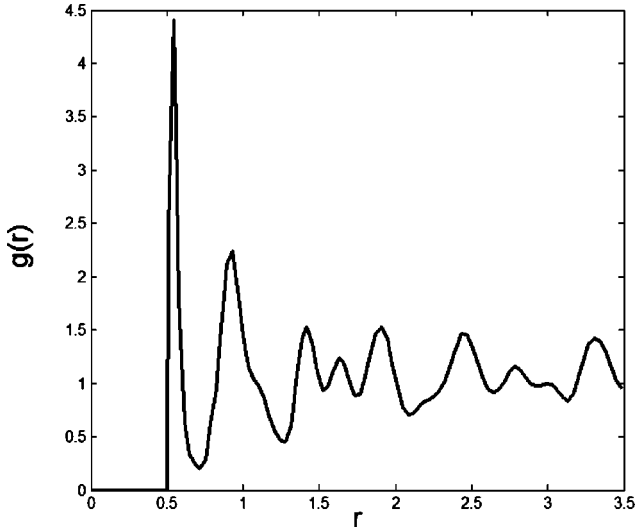


FIG. 7. Angle-averaged radial distribution function for hard triangles at high ($\rho=0.85$) packing fraction. Note the decay in successive maxima indicating a lack of long-range-translational order.

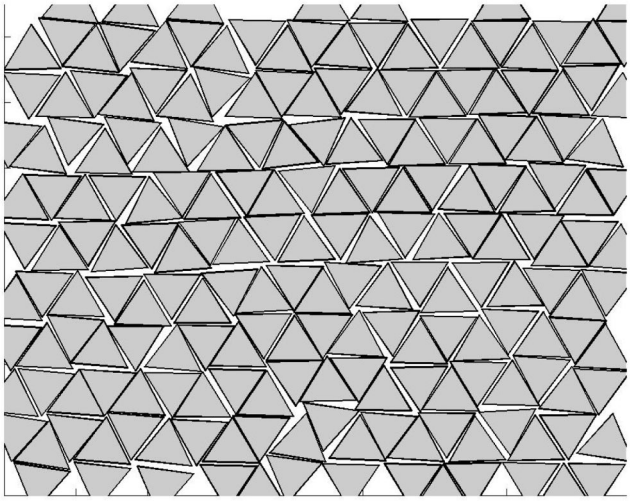


FIG. 8. Typical configuration at high packing fraction. There is perfect orientational order, but the triangles can move along slip lines.

correlated in their motion in that the center of mass of the two triangles is moving in a straight line. Collision with a third triangle is also of the chattering type, so that a system develops where there are extended correlations even at low density. This leads to configurations that are stable and long lived and in which the energy is equipartitioned and rigorously constant with time. Further work is underway to explore the virial coefficients and the equilibrium aspects of the system at low density.[20]

At high packing fraction, the radial distribution function, $g(r)$ has a number of well defined peaks (Fig. 7) that look somewhat solid-like, but note that they are relatively broad and decay with distance, indicating a lack of true long-range-translational order. Figure 8 shows a section of a typical equilibrated configuration at high packing fraction ($\rho=0.85$). The orientational order is nearly perfect ($g_6=.93$), but there is translational disorder due presumably to the ability to move along the slip planes. Around the transition point $\rho=0.63$, $g(r)$ takes on a structure that resembles a normal liquid (Fig. 9).

A configuration for this state is shown in Fig. 10. In this figure a few clusters have been highlighted to emphasize the tendency towards hexagonal clustering. Examination of movies of this state[20] show that such distorted clusters persist for times that are up to an order of magnitude longer than the mean time between collisions, even though g_6 has dropped to around 0.5. Clearly, the *local* excluded area can be minimized by nucleating and growing such clusters. When the cluster forms it may rotate relatively freely, and if a triangle attempts to escape from the cluster it stands a good chance of being struck on the external edge by the vertex of a triangle in the surrounding medium. In this way momentum transfer from the surrounding fluid stabilizes the cluster for relatively long periods of time. The situation is somewhat analogous to the depletion effect[21] that leads to an apparent attraction between large colloidal particles suspended in a molecular fluid. It should be emphasized that in the present case the association is between particles of the same size without solvent. The important point is that this phenomenon gives rise to long-lived clusters that are clearly associated even though there is no attractive force between the particles.

It would seem plausible that similar effect might operate even in spherical systems if the angle dependent three-body term is included in the analysis. The net effect of this term would be to stabilize the critical nucleus for times long enough to allow the attainment of a critical size.

The radial distribution function at low ($\rho=0.05$) packing fraction is shown in Fig. 11. Here each triangle can move on the average in an area that is twenty times greater than the intrinsic area of the particle. It might be thought that in this limit the ideal gas regime would prevail. This is not the case, and a curious phenomenon is observed. In the limit of very low packing fraction we expect, $g(r)=e^{-u(r)/kT}$, which for

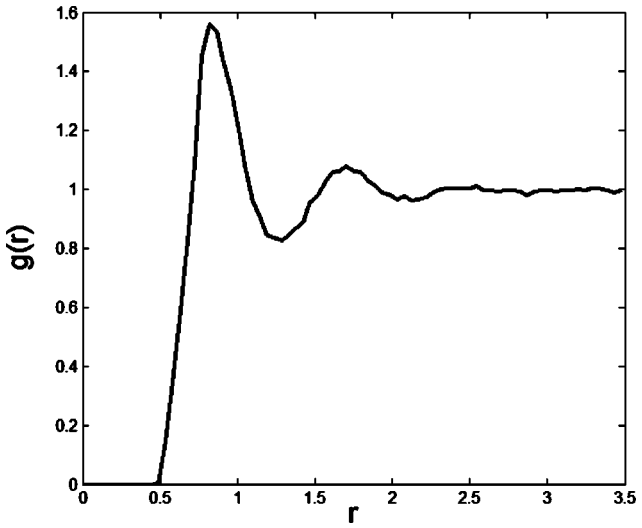


FIG. 9. Radial distribution function at intermediate packing fraction. Note the typical liquid-like profile with complete decay within ~3 "molecular" diameters.

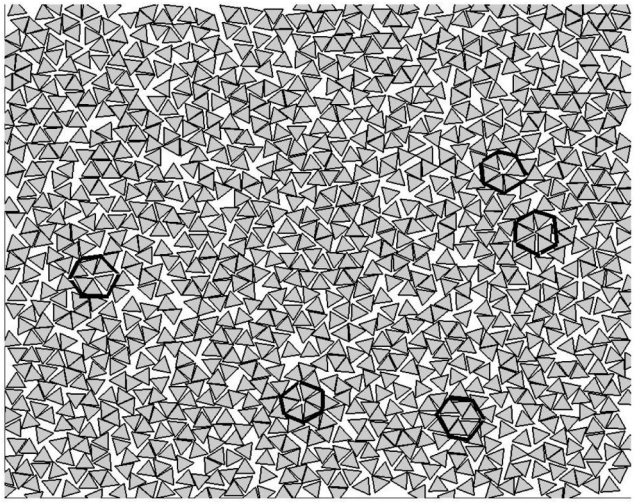


FIG. 10. Configuration in the neighborhood of the transition highlighting the tendency to form long-lived hexagonal clusters. Note also the visual similarity to a quasicrystal.

hard spheres is just the Heaviside step function. For triangles the situation is modified slightly, in that $g(r)$ rises from 0 to 1 over the distance range 0.5 to 1.0, reflecting the finite probability of collisions at these distances. However, examination of Fig. 11 shows a small excess probability at $r \sim 1.1$. This is a common feature of all of our simulations at low density. We have conducted a number of Monte Carlo simulations in this density regime. Running with very small angular and translational displacements for over 5×10^6 moves gave similar evidence of the small peak seen in the MD simulation.

A low density configuration is shown in Fig. 12. At short distance we can see clearly that the feature at $r \sim 1.1$ can be

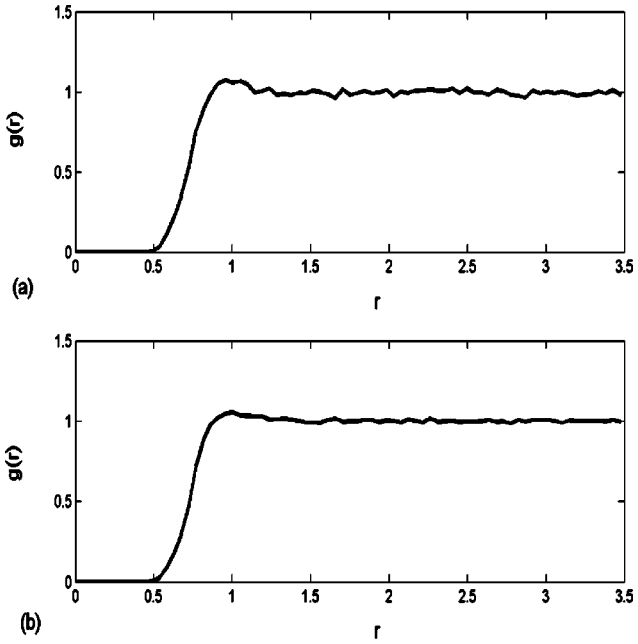


FIG. 11. (a) Radial distribution function for hard triangles at packing fraction 0.05. (b) $g(r)$ obtained from a Monte Carlo calculation at the same packing fraction.

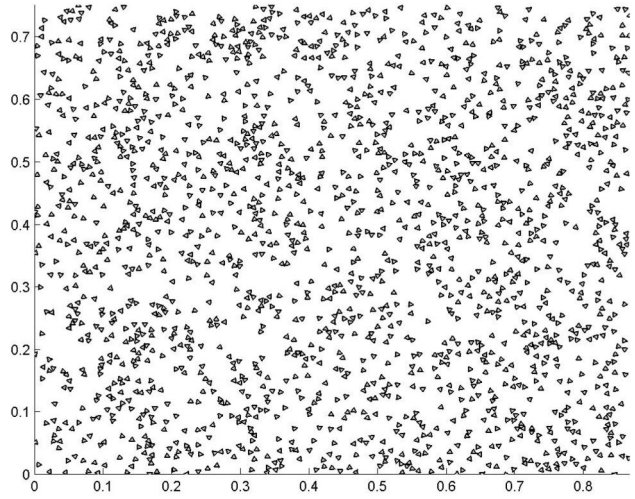


FIG. 12. Clustering of hard triangles at ~.1. Here each triangle has an average free volume that is an order of magnitude greater than the volume of the triangle. There is no attractive force between the particles and the clustering at low densities is due to the symmetry of the potential.

safely ascribed to triangles that are undergoing chattering collisions, and these events are not rare. There is also an apparent clustering effect at long distances. In the low density phase distorted hexamers are no longer present, but there is a clear tendency to form dimers, trimers, and other "linear" species. This is interesting because it means that the triangular symmetry of the potential without any attractive terms is sufficient to induce a kind of "bonding" with no attractive part in the potential. These particles stay together, not because they are attracted to each other, but rather because they cannot get out of each other's way.

In Fig. 13 the "effective" translational self-diffusion coefficient, calculated from Einstein's formula, is presented as a function of packing fraction. In general terms the behavior is comparable to that observed in spherical systems. Note, however, that the self-diffusion coefficient reaches a finite

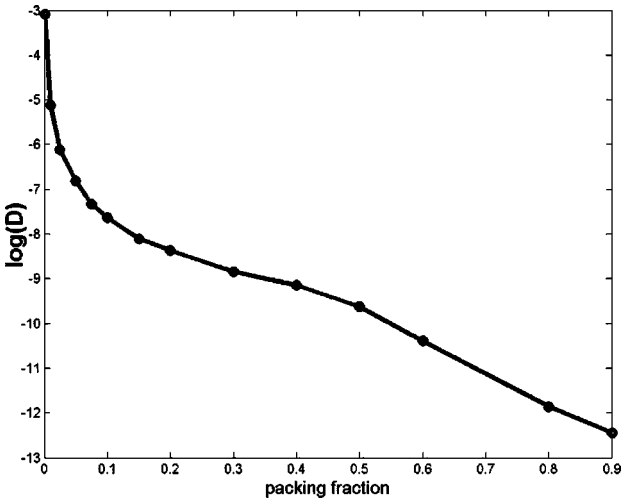


FIG. 13. The translational self-diffusion coefficient as a function of packing fraction.

limiting value of $\sim 10^{-13}$ as the packing fraction tends to unity.

V. CONCLUSIONS

This paper expands upon some preliminary results that were reported earlier in conference proceedings.[18,22] There is intense interest in granular systems as evidenced by the roughly 2000 recent citations listed in Ref. 23, but there is still no consensus as to the classical physics. Granular media are formally thermodynamically "small" systems that may not be in equilibrium and that exhibit dissipative interactions, leading Kadanoff[24] to speculate that attempts to describe granular materials using continuum approaches analogous to the Navier-Stokes formalism for liquids[25] may not be well founded. An ability to construct exact computer models in which the interparticle interactions, shape, and polydispersity are all well defined and from which the thermomechanical and transport coefficients can be calculated exactly is likely to be of value in a wide range of computational experiments.[26]

In summary, we have presented an approach to MD simulation of hard systems and have illustrated the method by applying it to the simplest example: the hard triangle in two dimensions. Even this simple example has revealed a surprising richness of phenomena that will require extension in fundamental theory ranging from the kinetic theory of gases to a better insight into the factors that are important in the nucleation and growth of crystals from an equilibrium fluid.

*Email address: john.maguire@wpafb.af.mil

[1] D. Frenkel and B. Smit, *Understanding Molecular Simulations* (Academic, New York, 1996).
[2] G. Ciccotti, D. Frenkel, and I. R. McDonald, *Simulation of Liquids and Solids* (Elsevier, Amsterdam, 1989).
[3] J. S. Rowlinson and F. L. Swinton, *Liquids and Liquid Mixtures*, 3rd ed. (Butterworths, London, 1982).
[4] J. S. Rowlinson and B. Widom, *Molecular Theory of Capillarity* (Oxford University Press, Oxford, 1982).
[5] D. Henderson, *Fundamentals of Inhomogeneous Fluids* (Marcel Dekker, New York, 1992).
[6] B. J. Alder and T. E. Wainright, J. Chem. Phys. **31**, 459 (1959).
[7] S. Auer and D. Frenkel, Nature (London) **413**, 711 (2001).
[8] C. Mao, V. R. Thalladi, D. B. Wolfe, S. Whitesides, and G. M. Whitesides, J. Am. Chem. Soc. **124**, 14508(2002).
[9] G. M. Whitesides and B. Grybowski, Science **295**, 2418 (2002).
[10] M. de Wild, S. Berner, H. Susuki, A. Bartoff, H. J. Guentheradt, and T. A. Jung, Proceedings IPMM, Sendai, Japan, 2003; M. de Wild *et al.*, Chem. Phys. 10, 881 (2002).
[11] D. Frenkel and J. F. Maguire, Mol. Phys. **49**, 503 (1983); Phys. Rev. Lett. **47**, 1025 (1981).
[12] J. J. Magda, H. T. Davies, and M. Tirrell, J. Chem. Phys. **85**, 6674 (1986).
[13] T. Aspelmeier, M. Huthmann, and A. Zeppelius, *Free Cooling of Particles with Rotational Degrees of Freedom*, in Granular Gases, edited by T. Poschel and S. Herding (Springer, Berlin, 2001).
[14] Y. Yoshimura and A. Mukoyama, J. Phys. A **34**, 4053 (2001).
[15] A. R. Barron, IEEE Trans. Inf. Theory **39**, (1993).
[16] D. W. Robertus and K. M. Sando, J. Chem. Phys. **67**, 2585 (1977).
[17] J. A. Anderson, *Neurocomputing: Foundations of Research* (MIT Press, Cambridge, MA, 1988).
[18] John F. Maguire, M. Benedict, S. LeClair, and L. Woodcock, Mater. Res. Soc. Symp. Proc. **700**, 241 (2002).
[19] A. Witkin, D. Baraff, and M. Kass, SIGGRAPH Course Notes (www.pixar.com), 1997.
[20] M. Benedict and J. F. Maguire (unpublished).
[21] See, for example, papers in Faraday Discuss. **112** (1999).
[22] J. Maguire, and M. Benedict, *A Solution to the N-Body Problem for Partiles of Non-Spherical Geometry Using Artificial Intelligence*, Proceedings of the Fourth International Conference on Intelligent Processing and Manufacturing of Materials (IPMM-03), Sendai, Japan, May, 2003.
[23] URL: www.ical1.uni.stuttgart.de/~lui/REFS/references.html
[24] L. P. Kadanoff, Rev. Mod. Phys. **71**, 435 (1999).
[25] C. Bizon, M. D. Shattuck, J. B. Swift, and H. L. Swinney, Phys. Rev. E **60**, 4340 (1999).
[26] R. Saksena, L. Woodcock, and J. Maguire, Mol. Phys. **102**, 259 (2004).

On the development of a confocal Rayleigh-Brillouin microscope

David C. Liptak, Jason C. Reber, and John F. Maguire[a)]
Air Force Research Laboratory, Materials Directorate, Wright-Patterson Air Force Base, Ohio, 45433-7750

Maher S. Amer
Air Force Research Laboratory, Materials Directorate, Wright-Patterson Air Force Base, Ohio 45433-7750 and Department of Mechanical and Materials Engineering, Wright State University, Dayton, Ohio, 45435

(Received 13 June 2006; accepted 11 December 2006; published online 31 January 2007)

This Note illustrates how a confocal microscope may be modified to conduct Rayleigh-Brillouin mapping experiments that yield very useful information on the mechanical properties of interfacial materials in small volume elements. While the modifications to the microscope are quite straightforward, they do entail significant changes in the optical design. The instrument described herein consists of an argon ion laser equipped with an actively stabilized intercavity étalon that serves as the excitation source for a modified Zeiss LSM 310 confocal laser scan microscope. The optics of the microscope were reconfigured to enable interfacing of the microscope with a tandem triple-pass Fabry-Pérot interferometer. This instrument enables three-dimensional Rayleigh-Brillouin spectral mapping of samples at micron spatial resolution. The performance of the instrument and its ability to perform both lateral and depth scans of the acoustic phonon velocity and, hence, the longitudinal modulus across bonded polymer/polymer and polymer/ceramic interfaces are illustrated and discussed. © *2007 American Institute of Physics.*
[DOI: 10.1063/1.2431181]

It is well known that the Rayleigh-Brillouin spectrum of a material contains a great deal of useful physicochemical and mechanical information.[1] For example, the width of the central Rayleigh line gives a measure of the thermal diffusivity of the material, the frequency shifts of the Brillouin lines give a measure of the adiabatic speed of sound that can be correlated to the longitudinal modulus, and the integrated intensity ratio of the Rayleigh and the Brillouin lines (known as the *Landau-Placzek ratio*) provides a measure of the heat capacity ratio of the material[2] and was also found useful in estimating the excess Gibbs free energy of mixing in liquid mixtures.[3] This feature could be important in that it may enable direct experimental measurements of free energy gradients at micron level spatial resolution within interfacial zones. It is particularly noteworthy that this wealth of information can be obtained using a light scattering technique that requires no physical contact with the material under investigation. The technique, therefore, is particularly well suited to making remote measurements on specimens that are small or that exist in inaccessible or hostile environments. In this Note, we describe the design, construction, and test of a confocal Rayleigh-Brillouin microscope with mapping capabilities. This instrument enables light scattering measurements to be made on micron-sized samples. Moreover, the instrument has confocal and mapping capability that enables depth profiling within samples and, most importantly, mapping across interfaces. This enables probing *local* properties of small volume elements that may be significantly different from macroscopic bulk properties of the material system.[4]

[a)]Author to whom correspondence should be addressed; electronic mail: john.maguire@wpafb.af.mil

A Spectra-Physics BeamLok 2060 argon ion laser fitted with an intercavity étalon for single frequency operation at (λ=514.5 nm) was used as excitation source. The laser is equipped with a dedicated controller that translates any laser frequency deviation into an error signal which drives a proportional change in the laser cavity length via a piezoelectric driven output coupler limiting the frequency jitter to less than 2 MHz over a 1 s time interval.[5] The dispersion and detection component is a triple tandem Fabry-Pérot interferometer manufactured by JRS Scientific Instruments.[6] The interferometer was equipped with two dynamic vibration isolation mounts that limit the frequency broadening due to étalon jitter to less than 1 kHz. For the imaging component, we used a Zeiss LSM 310 confocal laser scan microscope. The optical system of the confocal microscope was reconfigured such that the backscattered light of the Ar^+ laser could be imaged visually and directed to the input pinhole of the Fabry-Pérot interferometer. A schematic representation of the optical elements of the instrument is given in Fig. 1. A hole was drilled in the lower side of the LSM housing through which the argon laser was directed along the optical path of the microscope using a set of mirrors as shown in Fig. 1. The beam raster assembly of the LSM was removed. The beam is then directed using a beam splitter through the microscope objective and is focused onto the sample. The sample was mounted on an *XYZ* translation stage that has 1 μm spatial resolution in each direction. Backscattered light (180° scattering angle) is collected by the objective, collimated and passed in the reverse direction to a lens which focuses the light onto the confocal pinhole (PH1). On the far side of the pinhole the light is collimated once again before exiting the microscope through a second hole drilled in the upper side of the microscope.

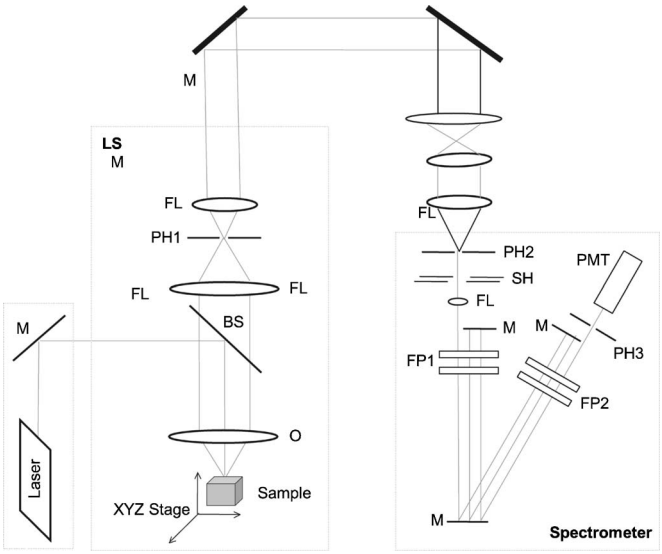


FIG. 1. Schematic diagram of optical components of the confocal Rayleigh-Brillouin microscope.

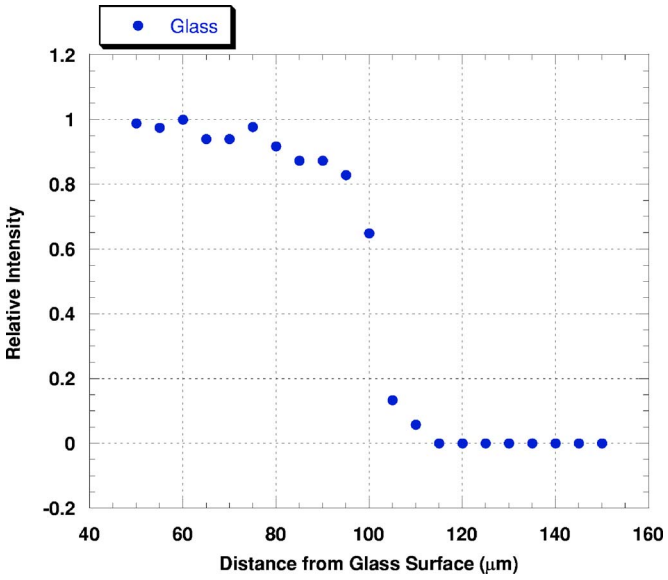


FIG. 2. Glass Brillouin peak intensity as a function of depth from the surface of a 100 μm glass slide on PMMA substrate.

The collimated beam is directed through a set of mirrors and lenses and then focused onto the spectrometer entrance slit using an achromatic lens [300 mm focal length 50.8 mm diameter] that matches the spectrometer $f/18$ requirement. Within the spectrometer the scattered light passes three times through the Fabry-Pérot étalon (FP1, FP2) before reaching the output pinhole (PH3) and finally the photomultiplier tube (PMT).

The performance of the instrument was evaluated by carrying out two linear Rayleigh-Brillouin mapping experiments. In the first experiment we scanned along a line in the xy plane across a polymethylmethacrylate (PMMA)/silicone rubber interface to demonstrate the instrument mapping capabilities and to determine the instrument lateral spatial resolution. In the second experiment, we scanned along a line in the z direction (depth profile scan) through a glass/silicone rubber/PMMA layered system to demonstrate the instrument depth scan capabilities across interfaces. Our purpose here is simply to report the performance of the spectrometer rather than to enter into any detailed discussion of the interfacial mechanical properties of multicomponent solid systems.

For these measurements the single spectral feature of interest is the Brillouin peak position of the longitudinal phonon mode as it relates directly to the longitudinal phonon velocity. Intensity of Brillouin peaks across the interface between two materials changed as shown in Fig. 2. For spatial resolution determination purposes, the cut-off limit was taken at 10% of the maximum intensity of the peak.

Lateral scan measurements were made on a sample having a PMMA and silicone rubber (Sylgard® 184 by Dow Corning) interface. Brillouin spectra were acquired from 32 consecutive points across the interface with a 1 μm step size. PMMA showed a Brillouin peak around 15.8 GHz, while the silicone rubber showed a peak around 6.1 GHz. The frequency shift ($\Delta\nu$) allows the calculation of the adiabatic speed of sound (C_s) in the material as follows:

$$C_s = \Delta\upsilon/q,$$

where q is the scattering vector with a magnitude given by[1]

$$|q| = \frac{4\pi n}{\lambda}\sin\left(\frac{\gamma}{2}\right),$$

where n is the refractive index of the probed material, λ is the wavelength of the incident light (514.5 nm), and γ is the scattering angle (180°).

From the measured phonon velocities, we calculated the longitudinal modulus according to

$$E = c_S^2\rho,$$

where ρ is the material's density. Using these equations and the auxiliary material constants shown in Table I (Ref. 7) with the measured phonon velocity as input, we may derive longitudinal modulus as a function of position across the interface as shown in Fig. 3.

The region where the two traces overlap shows the location and width (about 5 μm) of the interfacial region as resolved by our instrument. This overlap is much greater than the microscope optical lateral resolution of approximately 0.5 μm. This is not unexpected as the boundary between the two polymers is not discrete and, in fact, has a considerable degree of roughness at the edges. Therefore, the 5 μm width of the overlap region should not be considered as the best lateral resolution of the instrument.

Depth profiling measurements were performed on a sample made from a glass slide (approximately 100 μm thick) that was bonded to a 2 mm thick plate of PMMA by a commercial silicone based adhesive (Silicon II by GE). The thickness of the adhesive layer was approximately 30 μm. Twenty-six measurements were taken along the z axis down-

TABLE I. Material properties used in calculating the longitudinal moduli reported.

Material	n	ρ (g cm^3)	q (nm^{-1})
Glass	1.473	2.23	0.036
Silicone rubber	1.37	0.9548	0.0335
PMMA	1.48	1.19	0.0362

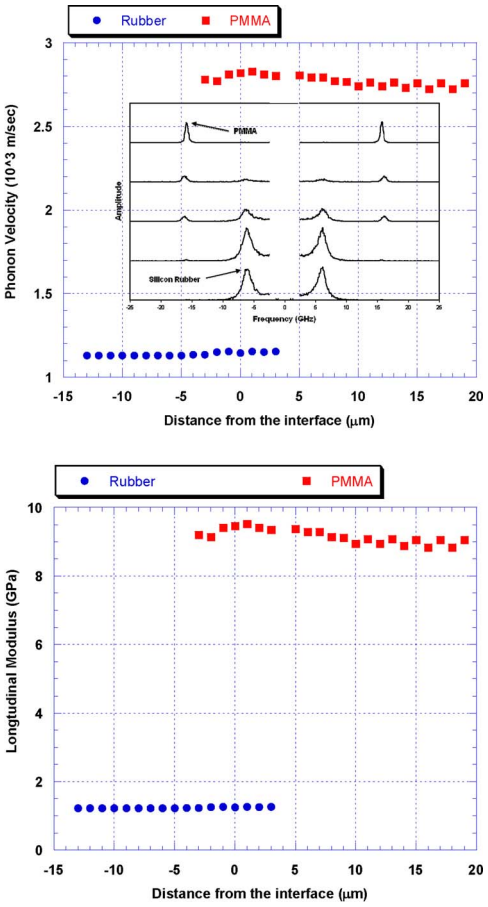


FIG. 3. (a) Phonon velocity vs position for lateral scan. (b) Calculated longitudinal modulus vs position for lateral scan.

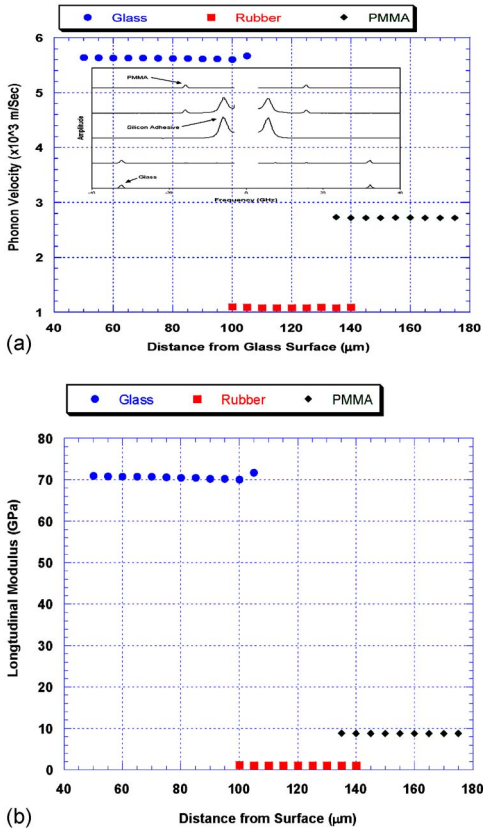


FIG. 4. (a) Phonon velocity vs position for depth scan. Insert, from bottom to top, Brillouin spectra of glass, silicon adhesive, and PMMA. (b) Calculated longitudinal modulus vs position for depth scan.

ward with a step size of 5 μm. The measured phonon velocity and the correspondingly calculated longitudinal modulus as a function of position along the depth scan are shown in Figs. 4. It also appears that the depth resolution of the instrument is approximately 5 μm.

While Koshi and Yarger have recently noted some of the significant benefits and potential power of this form of spectroscopy,[8] and have reported Brillouin images of liquid and polymer materials, their experimental resolution of the images was 20 μm in the lateral and over 500 μm in the depth direction. It is noteworthy that our instrument enables significantly higher spatial resolution by an order of magnitude in the lateral direction and over two orders of magnitude in the depth direction. Moreover, it allows visual observation of the region under study.

The utility of this instrument was demonstrated by creating lateral and depth maps of phonon velocities and longitudinal modulus for layered material systems. The instrument proved capable of resolving material properties with a spatial resolution in the range of 5 μm.

[1] B. Berne and R. Pecora, *Dynamic Light Scattering* (Dover, New York, 2000).
[2] L. Landau and G. Placzek, Phys. Z. Sowjetunion **5**, 172 (1934).
[3] J. F. Maguire, J. C. F. Michielson, and G. Rakhorst, J. Phys. Chem. **85**, 2972 (1981).
[4] D. Mehtani, N. Lee, R. D. Hantschuh, A. Kisluik, M. D. Foster, A. P. Sokolov, and J. F. Maguire, J. Raman Spectrosc. **36**, 1068 (2005).
[5] *User's Manual: Model 588 J-Lok® with Model 5880 Controller* (Spectra-Physics, San Diego, CA, 1998).
[6] S. Lindsay, M. Anderson, and J. R. Sandercock, Rev. Sci. Instrum. **52**, 1478 (1981).
[7] *CRC Handbook of Chemistry and Physics*, 87th ed. (CRC Press, Boca Raton, 2006).
[8] K. Koski and J. Yarger, Appl. Phys. Lett. **87**, 061903 (2005).