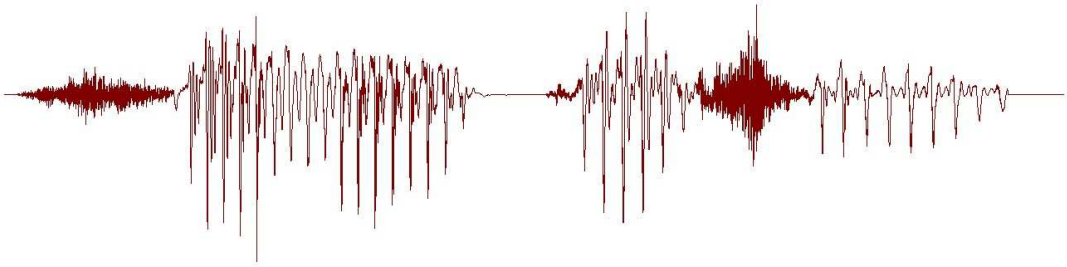
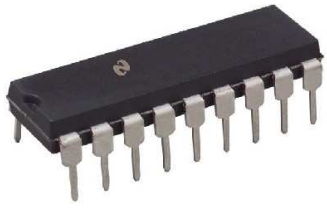
the

Phonetician

A publication of ISPhS/International Society of Phonetic Sciences





Number 97/98 2008, I - II

ISPhS
International Society of Phonetic Sciences

President: Ruth Huntley Bahr

Secretary General:
Mária Gósy

Honorary President:
Harry Hollien

Executive Vice President: Eric Keller

Vice Presidents:
Angelika Braun
Maria Dohalská-Zichová
Mária Gósy
Damir Horga
Eric Keller
Heinrich Kelz
Stephen Lambacher
Asher Laufer
Judith Rosenhouse

Past Presidents:
Jens-Peter Köster
Harry Hollien
William A. Sakow †
Martin Kloster-Jensen
Milan Romportl †
Bertil Malmberg †
Eberhard Zwirner †
Daniel Jones †

Honorary Vice Presidents:

A. Abramson	P. Janota	P. Ladefoged †	R. K. Potapova	F. Weingartner
S. Agrawal	W. Jassem	A. Marchal	M. Rossi	R. Weiss
L. Bondarko	M. Kloster-Jensen	H. Morioka	M. Shirt	
E. Emerit	M. Kohno	R. Nasr	E. Stock	
G. Fant	E.-M. Krech	T. Nikolayeva	M. Tatham	

Regional Secretaries:

M. Bakalla	Mid-East
M. Dohalská-Zichová & T. Duběda	Czech Section
C. Paboudjian	Southern Europe
R. K. Potapova	Russia
P. French	United Kingdom
Y. H. Wang	South East Pacific
H. Yamazawa	Japan

Auditor:
Angelika Braun

Treasurer:
Ruth Huntley Bahr

Affiliated Members (Associations):

American Association of Phonetic Sciences	J. Hoit & W. S. Brown
Dutch Society of Phonetics	B. Schouten
International Association for Forensic Phonetics and Acoustics	A. Braun
Phonetic Society of Japan	I. Oshima & K. Maekawa
Polish Phonetics Association	G. Demenko

Affiliated Members (Institutes and Companies):

Kay Elemetrics, Lincoln Park, NJ, USA	J. Crump
Inst. for Advanced Study of the Communication Processes, Univ. Florida, USA	H. Hollien
Dept. of Phonetics, University of Trier, Germany	J.-P. Köster
Dept. of Phonetics, University of Helsinki, Finland	A. Iivonen
Dept. of Phonetics, University of Zürich, Switzerland	S. Schmid
Centre of Poetics and Phonetics, University of Geneva, Switzerland	S. Vater

INTERNATIONAL SOCIETY OF PHONETIC SCIENCES (ISPHS): ADDRESSES
www.isphs.org

President:
Prof. Ruth Huntley Bahr, Ph.D.
President's Office:
University of South Florida
Dept. of Communication Science and Disorders
4202 E. Fowler Ave., PCD 1017
Tampa, FL 33620-8200
USA
Tel.: ++1-813-974-3182
Fax: ++1-813-974-0822
e-mail: rbahr@chuma1.cas.usf.edu

Secretary General:
Prof. Mária Gósy, DSc
Secretary General's Office:
Kempelen Farkas Speech Research Laboratory
Hungarian Academy of Sciences
Benczúr u. 33
H-1068 Budapest
Hungary
Tel.: ++36 (1) 321-4830 ext. 172
Fax: ++36 (1) 322-9297
e-mail: gosy@nytud.hu

Guest Editors:
Prof. Gábor Olaszy, DSc
and
Prof. Géza Németh, Ph.D.
Guest Editors' Office:
Department of Telecommunications and Mediainformatics
Budapest University of Technology and Economics
Magyar tudósok krt. 2
H-1117 Budapest
Hungary
Tel.: ++36 (1) 463-3883
Fax: ++36 (1) 463-3107
e-mail: olaszy@tmit.bme.hu
nemeth@tmit.bme.hu

Review Editor:
Prof. Judith Rosenhouse, Ph.D.
Review Editor's Office:
Swantech
89 Hagalil St
Haifa 32684
Israel
Tel.: ++972-4-8235546
Fax: ++972-4-8327399
e-mail: swantech@013.net.il

the
Phonetician

A Publication of ISPhS/**International Society of Phonetic Sciences**

ISSN 0741-6164

Number 97 / 2008-I-II

CONTENTS

From the President **6**

Papers **7**

A formant trajectory database of Hungarian vowels *by Gábor Olaszy, Zsuzsanna Zsófia Rácz and Kálmán Abari* 7

A comparative study of direct and ASR-based modular audio to visual speech systems *by Gergely Feldhoffer, Attila Tihanyi and Balázs Oroszi* 15

Improving naturalness of visual speech synthesis *by László Czap and János Mátyás* 27

Exploring differences between phonetic classes in Sleep Apnoea Syndrome Patients using automatic speech processing techniques *by Jose Luis Blanco, Rubén Fernández, Eduardo López and Luis Alfonso Hernández* 36

Phonetics Institutes Present Themselves **56**

Laboratories of Speech Research & Technology at the Department of Telecommunications and Media Informatics (TMIT), Budapest University of Technology and Economics (BME) *by Géza Németh* 56

Ph.D. research **60**

Integrating prosody into Automatic Speech Recognition *by György Szaszák* 60

Research project **68**

Report on building a tool for Romanian spontaneous speech recognition *by Corneliu Burileanu, Cristina-Sorina Petrea, Andi Buzo, Horia Cucu and Alina Pasca* 68

Tools using speech technology for research and education **83**

Glottalizer: A tool to transform regular voice into irregular voice *by Tamás Bőhm and Nicolas Audibert* 83

On line acoustic features of segmental level speech data (Hungarian) *by Gábor Olaszy* 88

From text-to-sound - A Hungarian word level pronunciation database using IPA symbols *by Kálmán Abari* 93

Applications using speech technology research results 105

Mindroom – Video news search system by voice *by Tibor Fegyó and Sándor Somos* 105

Spoken dialogue-based phone information system for pharmaceuticals *by Gábor Olaszy* 109

On the Hungarian Language and Speech Technology Platform *by Tamás Váradi* 119

Book reviews 122

Lingua Americana Revista de Linguistica *reviewed by Judith Rosenhouse* 122

Sánchez Miret, Fernando: La diptongación en las lenguas románicas *reviewed by José María García Martín* 124

Daniel Jones (edited by Peter Roach, James Hartman and Jane Setter): Cambridge English Pronouncing Dictionary *reviewed by Wiktor Jassem* 130

Watson, Janet C.E.: The phonology and morphology of Arabic *reviewed by Judith Rosenhouse* 134

Richard V. Teschner & M. Stanley Whitley: Pronouncing English *reviewed by Wiktor Jassem* 138

Daniel Schreier: Consonant change in English worldwide. Synchrony meets diachrony *reviewed by Judith Rosenhouse* 141

Jeroen van de Weijer, Kensuke Nanjo, Tetsuo Nishihara (eds.): Voicing in Japanese *reviewed by Eric Rosen* 144

Gábor Olaszy: Mássalhangzó-kapcsolódások a magyar beszédben [Consonant Clusters in Hungarian Speech] *reviewed by Péter Siptár* 150

Hangidőtartamok és időszerkezeti elemek a magyar beszédben [Sound durations and temporal factors in Hungarian speech] *reviewed by Kálmán Abari* 152

Jacob Benesty, M. M. Sondhi and Yiteng Huang (eds.): Handbook of speech processing *reviewed by Eduardo López Gonsalo* 154

Meetings, conferences and workshops 156

Call for papers 176

Instructions for book reviewers 176

ISPhS Membership Application Form 177

News on Dues 178

FROM THE PRESIDENT

A new double issue of the *Phonetician* is now ready. Our guest editors, Géza Németh and Gábor Olaszy, have done an excellent job! I am extremely grateful for their hard work in putting this issue together. They have brought together a fine group of researchers, who have shared their expertise with us in the area of speech technology. This issue offers both traditional research papers, as well as articles describing practical applications of speech technology research. There is even a section on tools that involve speech technology. I particularly enjoyed reading about the applications involving medicine, voice analysis, and pronunciation databases. This issue reminds the reader of the importance of speech technology in our everyday lives. In addition, it includes book reviews and a listing of meetings, conferences and workshops. These features are the backbone of any issue of the *Phonetician*. I hope you enjoy reading it as much as I did.

As usual, we are looking for members to take a more active role in the production of the *Phonetician*. We need to know more about *your* area of specialization. I know that you have colleagues and peers that are making a difference in the area of phonetics. We want to hear about that. With so many Phonetics programs struggling to survive in the midst of academic budget woes, researchers in the phonetic sciences need to stick together. We need each other now more than ever. The officers in ISPhS do recognize that peer-reviewed journals are essential for the exchange of ideas and research findings. As scientists, we value the feedback we receive from our peers. It calls us higher and helps us generate new thoughts and ideas. But there is also a need for just sharing ideas and learning about what other researchers are doing. ISPhS strives to fill that need. We want to bring researchers in the phonetic sciences together and support one another in research projects. The *Phonetician* is a particularly good place for your students to present their work to an international audience.

So, what have YOU contributed to the *Phonetician?* Have you sent in a report describing your work or your laboratory? Have you reviewed a book for us lately? I know that Prof. Rosenhouse has a book for you to review, if you just offer. Did you write up a brief description of the last professional meeting sponsored by your institution? These types of contributions are essential for the continuation of the *Phonetician*. We need YOU to take an active role. You can guest edit an entire issue (with our help) or you can submit an individual paper, laboratory description or book review to us to include in an upcoming issue of the *Phonetician*. Please keep the *Phonetician* alive. Just let me know what you would like to contribute!

A FORMANT TRAJECTORY DATABASE OF HUNGARIAN VOWELS

Gábor Olaszy*, Zsuzsanna Zsófia Rácz* and Kálmán Abari**
***Department of Telecommunications and Media Informatics**
Budapest University of Technology and Economics, Hungary
****University of Debrecen, Hungary**
e-mail: olaszy@tmit.bme.hu, zsuzska.racz@gmail.com, abari.kalman@gmail.com

Abstract

Previously, the investigation of the formant structure of Hungarian vowels involved individual measurements. However, the progress of speech science towards statistical analysis makes public databases necessary. In this paper, we report on the creation of such a formant database of Hungarian vowels, based on a speech corpus of isolated words read by both a male and a female speaker. The steps of its development are explained, from the automation of measurements to the manual data correction phases.

Possibilities for querying the database include formant maps, formant ranges and trajectories based on three measurement points per vowel. The accuracy of our formant measurement is also discussed.

This database can be used as a reference to facilitate linguistic studies, classroom measurements and individual research topics in university education.

1 Introduction

Formant measurement is one of the oldest areas in speech research. It is a matter of great importance both in linguistics and in automatic speech processing. Formant trajectories aid human perception, as well as the computer processing of speech.

In earlier Hungarian speech research, Tarnóczy (1941) was the first to determine the formant values of Hungarian vowels. He used an oscilloscope and a camera to analyse the wave forms and from these he calculated the Fourier coefficients and the formant ranges. Scientists studying the subject after him, Magdics (1965), Szende (1973), Bolla (1978), Olaszy (1989) and Gósy (2004) were able to utilize more modern equipment for measurement.

Previously, research studying the formant structure of the Hungarian language was based on individual measurements. In other words, the sound recordings used were not accessible for other experiments. Only the conclusions and not the raw formant data were published. Since speech science tends towards statistical analysis, public data are needed. The creation of formant databases is quite unique as a concept. We only know about one such database (Deng Li et al. 2006), containing formants of several speakers and dialects of English.

In this paper, we report on the creation of a Hungarian reference formant trajectory database. Measurements are based on recordings of isolated words, as opposed to fluent speech.

2 Corpus

The public corpus used in the project was created in 2006 (Abari & Olaszy, 2007; http://fonetika.nytud.hu/cvvc). It consists of 2912 recordings of approximately 1500 Hungarian words read in Standard Hungarian by a male and a female reader, ages 60 and 30 years respectively. The length of this material is 49 minutes. The corpus was recorded in an anechoic chamber at 16 bits and a sampling rate of 22 kHz.

The recordings were manually labelled and segmented at the phonetic level. Their phonetic transcription was completed by an automatic method and then checked and corrected by hand.

3 Development

From the several computerised formant estimation methods available, we selected the algorithm realised in Praat for our project. Praat is widely known and can be programmed using a simple scripting language which makes processing large data sets possible. (There are alternative ways of formant detection.)

The raw measurements were carried out using a Praat script capable of automatically processing the input, which was a set of large file folders containing waveforms with corresponding segmentation, label and phonetic transcription data. The script uses the formant function in Praat to calculate the first four formant frequencies of every vowel at certain moments, instead of recording continuous trajectories to represent the effect of the CVC (consonant – vowel – consonant) groups. The following measuring points have been selected: the 25%, 50% and 75% points of vowel duration to represent the centre of the vowel and both transient phases when the vowel was in an intermediate position; the 50% and 75% points; when it was at the beginning of the word and 25%, 50% points when it was at the end of the word, as we have found the spectrum of peripheric periods uncertain. The parameters of calculation (e.g., a window length of 25 ms) were the same as the defaults in Praat with the exception of the maximum formant, where 5000 Hz and 5500 Hz were used for recordings of the male and the female speaker, respectively.

The script generates a file suitable for importing into a spreadsheet or database. Each output file contains the following information: the sex of the reader; the phonetic description of the word; the vowel along with the neighbouring sounds (usually consonants); the position of the measurement points and the F_1-F_4 formant frequencies in Hertz.

The output of the script was a raw formant database containing 29,926 measurement points and 119,704 total formant values. The raw data contained errors because of uncertainties in pronunciation, inaccurate segmentation and/or signal processing faults. These errors had to be traced down and corrected.

Therefore, we used a spreadsheet application and a statistical analyser to filter out measurement errors. Criteria taken into account included formant ranges, formant distances and formant dependences (e.g. F_1/F_2 and F_3/F_4 maps). The filtering could only be partly automated, so a large part of it was completed manually. Inaccuracies were then corrected by visual observation of the spectrograms which can be found on the above mentioned website.

In the first step of data filtering, the measured formant frequencies were ordered by value, for every gender-vowel pair separately. This made it possible to find crude errors based on the difference between the measurements and the expected value. We determined the expected values based on *a priori* estimations and previous findings for Hungarian (see above).

Afterwards, the inspection of formant distances revealed further inaccuracies. For example, the expected minimum of F_2–F_1 is $2F_0$ (approximately 160 Hz for a male voice), so a record with an F_2–F_1 lower than this was deemed incorrect. The most extreme case in this respect was a 7 Hz difference detected between F_3 and F_2, but overall the number of records under 100 Hz was 25 for F_2–F_1, 127 for F_3–F_2 and 58 for F_4-F_3 (Olaszy et al., 2009). The fact that Praat distinguished these frequencies as formants raises questions about the built-in criteria in its estimation algorithm.

In the next step, the data obtained for the three measuring points were examined as a group. The reason for this was that these points represent the formant trajectories, which are typical of the CVC groups. For example, nasal consonants next to the vowel smear the formants due to coarticulatory nasalisation (the resonances of the nasal cavity smooth the formants of the vowel).

The final step of correction included inspection of the F_1-F_2 and F_3-F_4 maps for both speakers and for all vowels, which sped up the filtering of the previously undetected inaccuracies.

After the correction phase, the reference formant database was complete, that is, the number of erroneous data records in the database has been minimised.

4 Measurement results and options of use

Table 1 shows the structure of the reference database using an example word. All of the formant frequencies are given in Hz in the table below. No data for the 25% position for the very first vowel [aː] is present due to weak spectral data. The missing absolute word beginning (and ending position) is represented with a hash mark (‘#’) symbol.

The characteristic formant values for the first vowel are 838, 1418, 2743 and 3333 Hz. The transient phase to the dental-alveolar articulatory position of the neighbouring plosive is reflected by the data at the 75% point. The formant movements were: F_1 (down), F_2 (up), F_3 (down), F_4 (up). In the second vowel, [u], F_1 and F_4 do not show much movement in the transient phase. F_2 shows a continuous rise when going towards the following dental-alveolar trill, while F_3 a continuous decline across the sound. The third vowel [i] does not show formant movements between the dental-alveolar consonants.

Table 1. Data records in the reference formant database belonging to the word *átgurít* ([a:dguri:t], meaning ‘to roll something over’) read by the male speaker

Gender	Word	Previous	V	Next	Position	F_1	F_2	F_3	F_4
male	a:dguri:t	#	a:	d	50%	838	1418	2743	3333
male	a:dguri:t	#	a:	d	75%	702	1514	2662	3466
male	a:dguri:t	g	u	r	25%	335	849	2666	3618
male	a:dguri:t	g	u	r	50%	345	910	2629	3631
male	a:dguri:t	g	u	r	75%	336	1093	2472	3643
male	a:dguri:t	r	i:	t	25%	326	2160	2685	3584
male	a:dguri:t	r	i:	t	50%	313	2221	2706	3601
male	a:dguri:t	r	i:	t	75%	326	2217	2646	3665

In short, the database gives the main formant data for the 50% points and the formant movements inside the vowel as a function of adjacent sounds. This way, characteristic parameters can be compared through sound environment, articulatory place, or formant data.

For example, the distribution of formants can be obtained. Figure 1 shows the range of formant frequencies for all vowels, without taking the sound environment into account.

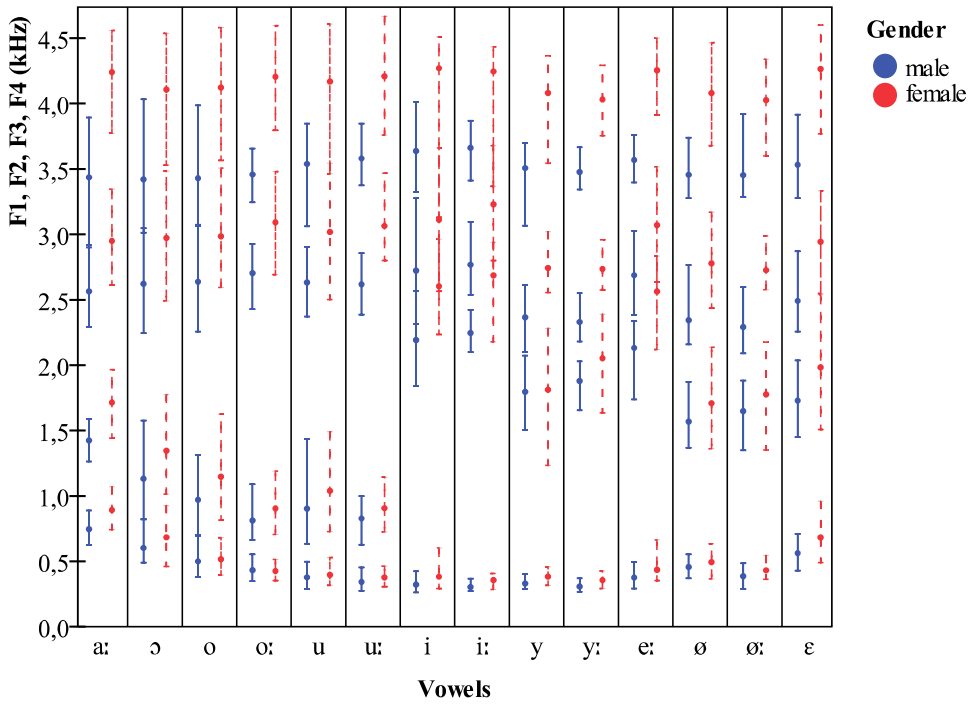


Figure 1. F_1-F_4 ranges of Hungarian vowels, based on the 50% point data in the reference database (ranges on the right correspond to the female speaker)

If we restrict our search by articulatory place, the distribution will be different. For example, the result of an analysis of vowels between dental-alveolar consonants is shown in Figure 2.

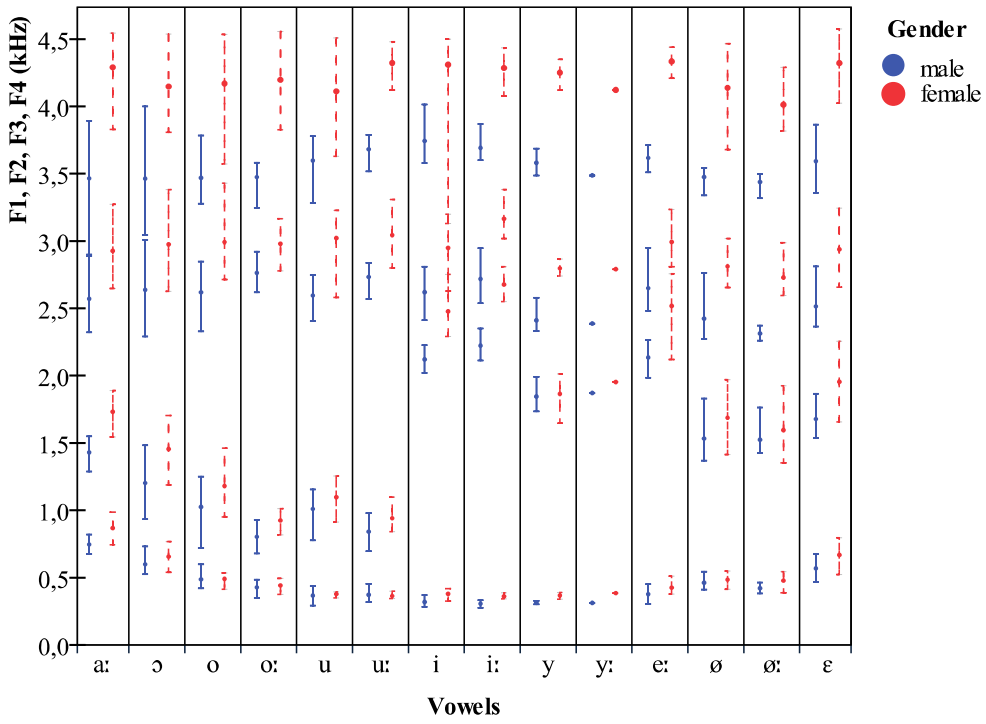


Figure 2. F_1-F_4 ranges of dental-alveolar – vowel – dental-alveolar clusters, based on the 50% point data in the reference database (ranges on the right correspond to the female speaker)

A third option of use of the database is shown in Figure 3, for which all the formant data of the male speaker have been taken into account, although by defining thresholds, it is possible to filter out outliers.

Formant trajectory graphs may also be generated. An example of this is shown in Figure 4. Palatal adjacent consonants raise the second formant, but lower the third, while F_1 remains constant. These formant movements give information about the locus (Delattre et al. 1955) of the palatal consonants. However, a labiodental consonant in the sound grouping would have an effect on the first formant as well.

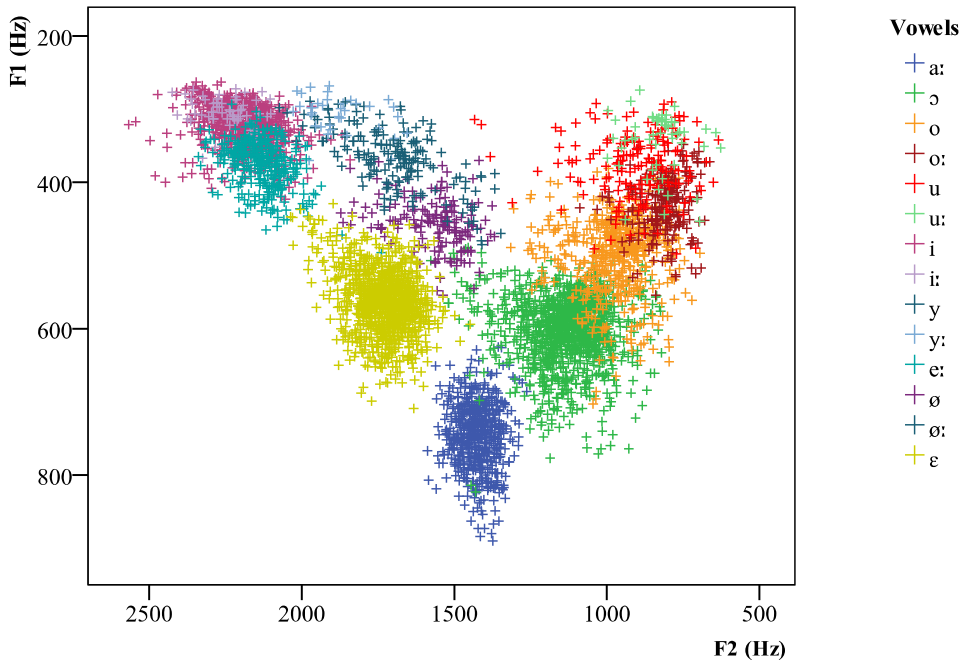


Figure 3. F_1-F_2 map of the central measurement points of Hungarian vowels for the male speaker, based on the reference database

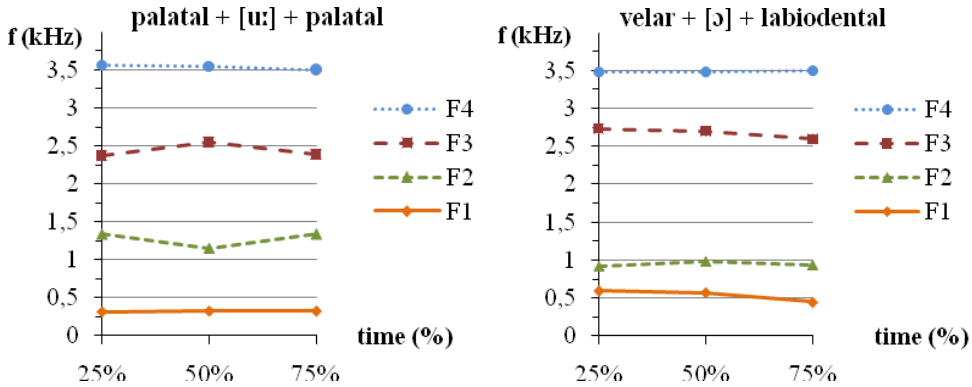


Figure 4. Average interpolated formant trajectories of vowels in two clusters of the male speaker (between two palatal consonants, and between a velar and a labiodental consonant)

5 The efficiency of formant estimation

The correction process of the raw database revealed the error ratios shown in Figure 5. A total of 10,307 formant values were corrected, 916 of which belong to the male reader. A total of 25.1% of all vowels had to be modified in at least one of the measured formants. The higher the formant, the more inaccurate the estimation

algorithm proved to be. A significant percentage of errors were due to the fact that Praat has placed formants too close to one another (that is, detected formants where there were none) or it failed to detect an existing formant.

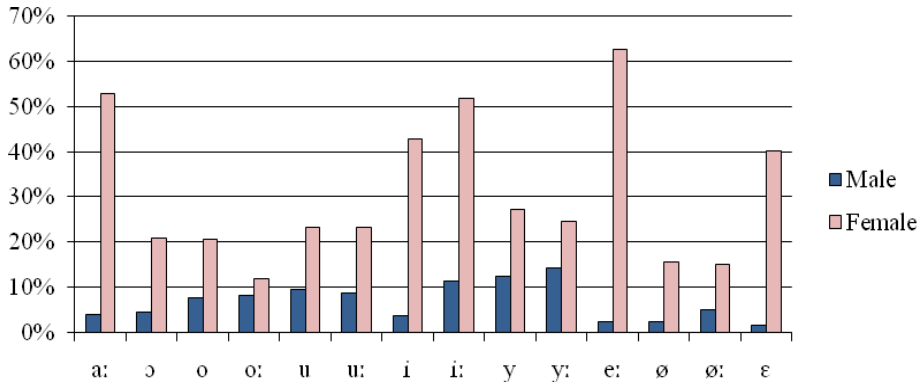


Figure 5. The rates of inaccurate formant measurements by Praat as a function of the vowels, for both speakers

Inaccuracy in higher formant measurement was particularly high in the case of the female speaker, due to the higher fundamental frequency of women. This means that the spectral lines of the quasi-periodic voiced sounds are further from each other than a male speaker's. Furthermore, the intensity of formants decreases with the frequency. Therefore, detecting the points of maximum amplitude correctly is more difficult for F_3 or F_4 than for F_1.

The phenomena leading to these inaccuracies are not specific to Praat, but are typical of formant measurement in general.

Figure 6 shows the error rates as a function of the formants. Inaccuracy for F_3 and F_4 of the female speaker was 21% and 31%, respectively.

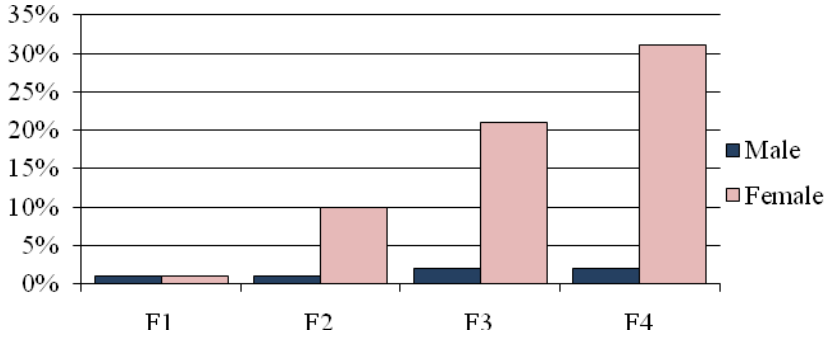


Figure 6. The rates of inaccuracy as a function of the formants, for both speakers

6 Conclusion

The aim of this paper was to describe a high precision Hungarian formant database which could be considered as a reference. We have explained the steps of development and given some options for use. Apart from the inspection of average formant values and ranges, these options also include examination of how the sound group affects formants frequencies (e.g., palatal consonants raise the frequency of the second formant). Measurements comparing genders or different vowels can also be carried out.

In addition, the database can be used for determining and increasing the accuracy of formant measuring algorithms, or for aiding speech research in other ways. It is important to note that the error rates shown in this paper have not been compared to those of other formant detectors and most of the inaccuracies noted in this project do not appear to be software-specific.

We hope to make the database available online.

References

Abari, K. and Olaszy, G. 2007. A magyar beszéd hangkapcsolódásainak bemutatása az interneten. *Beszédkutatás 2007*. 178-186.

Bolla, K. 1978. A magyar magánhangzók akusztikai analízise és szintézise. *Magyar Fonetikai Füzetek*, 1: 53-67.

Delattre, P. C., Liberman, A. M. and Cooper, F. S. 1955. Acoustic loci and transitional cues for consonants. *Journal of Acoustic Society America,* 27(4): 769–773.

Deng Li et al. 2006. A database of vocal tract resonance trajectories for research in speech processing. In *Proceedings of the ICASSP 2006.* 369-372.

Magdics, K. 1965. Magyar beszédhangok akusztikai szerkezete. *Nyelvtudományi értekezések 49.*

Olaszy, G. 1989. *Elektronikus beszédelőállítás.* Budapest: Műszaki Kiadó.

Olaszy, G., Rácz, Zs. Zs. and Bartalis, M. 2009. Formánsmérések automatizálása, formánsadatbázisok létrehozása. *Beszédkutatás 2009.* 134-147.

Szende, T. 1973. Spontán beszédanyag gyakorisági mutatói. *Nyelvtudományi értekezések 81*.

A COMPARATIVE STUDY OF DIRECT AND ASR-BASED MODULAR AUDIO TO VISUAL SPEECH SYSTEMS

Gergely Feldhoffer, Attila Tihanyi and Balázs Oroszi
Pázmány Péter Catholic University, Hungary
e-mail: flugi@itk.ppke.hu, tihanyia@digitus.itk.ppke.hu, oroba@digitus.itk.ppke.hu

Abstract

A comparative study of audio-to-visual speech conversion is described in this paper. A direct feature-based conversion system is compared to various indirect ASR-based solutions. These methods have been tested in the same environment in terms of audio pre-processing and facial motion visualization. Subjective opinion scores show that with respect to naturalness, direct conversion performs well. Conversely, with respect to intelligibility, ASR-based systems perform better.

1 Introduction

The goal of an audio-to-visual speech (ATVS) conversion system is to convert acoustic speech into visual speech. Such systems usually are comprised of an audio pre-processing component, audio-to-video (AV) mapping, a face model and a rendering subsystem. This paper will focus on methods used for performing the mapping from audio to visual speech.

There are different strategies for performing audio to visual conversion. One approach is to utilize automatic speech recognition (ASR) to extract phonetic information from the acoustic signal. This is then used, in conjunction with a set of coarticulation rules, to interpolate a visemic representation of the phonemes (Beskow et al 2004, Moubayed et al. 2008). Alternatively, a second approach is to extract features from the acoustic signal and convert these features directly to visual speech (Takács et al. 2006, Hofer et al. 2008).

A difficulty that arises in comparing the different approaches is they usually are developed and tested independently by the respective research groups. Different metrics are used, e.g. intelligibility tests and/or opinion scores, and different data and viewers are used (Theobald et al. 2008). In this paper, we describe a comparative evaluation of different AV mapping approaches within the same workflow see (Figure 1). The performance of each is measured in terms of intelligibility, where lip-readability is measured, and naturalness, where a comparison with real visual speech is made.

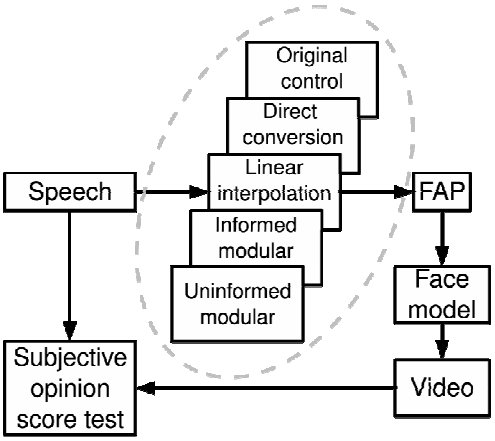


Figure 1. Multiple conversion methods were tested in the same environment

2 Related work

Our group started to work on speech audio visualization in 2004 with the goal of aiding deaf and hearing impaired people with mobile devices. Soon, the specification included audio-to-visual speech conversion. Because of the limited computational power of mobile devices, we focused on direct conversion, thereby avoiding phoneme or viseme classification tasks. The first working system was measured by word recognition test with deaf individuals, which resulted in a 48% recognition rate when compared to a real face. This result was achieved with a professional lip-speaker providing the training data. A database using an everyday speaker was not as good, with the recognition rate being substantially better than chance.

Speaker dependency was then investigated. This can be decreased by using additional audio data in the database, which is aligned with the main recording. This alignment can be automated using DTW (Dynamic Time Warping). This is important since professional lip-speakers are typically women, and we have to support male voice as well (Feldhoffer 2007).

Direct conversion uses time window of audio as input for the machine learning subsystem. We investigated the required size of this window using mutual information estimation between the auditory and visual modalities. An additional result was the temporal asymmetry of the connection, we found that the audio data could be partly predicted using video data (Feldhoffer et al. 2007).

2.1 Data capture and processing

To train the synthesis systems, a corpus of audiovisual speech spoken by a professional lip-speaker was recorded. The material included keywords to use as reference data, and 100 sentences, balanced for phonetic content and recorded with six speakers. The number of sentences varied between trainings. The database described in this paper included one speaker and only 3 sentences and the keywords. The acoustic speech was recorded in quiet and was sampled at 48 kHz, 16 bit mono.

The accompanying video was captured at 25 frames per second. The speakers was instucted to have a neutral expression. The head of the speaker was fixed by adjustable pillows to eliminate basic head motion, and a reference marker was placed on the nose. Lighting was adjusted to the marker tracking system and varied from speaker to speaker.

The speaker wore face markers, which were tracked in 2D pixel space. Any tracking errors were corrected manually, and the marker positions were median filtered to remove noise. In accordance with the MPEG-4 standard, the facial units (FAPU) of the model were calculated in this pixel space and the facial parameters were coded in FAP format (Facial Animation Parameters) using this FAPU. Principal components analysis (PCA), was then used to compress the resulting facial motions, where 6 coefficients were retained.

The acoustic speech was parameterized as 16-16 MFCCs of five windows.

3 Audio-to-visual conversion

The performance of five different approaches will be evaluated. These are summarized as follows:

- A reference based on natural facial motion.
- A direct conversion system.
- An ASR based system that linearly interpolates phonemic/visemic targets.
- An *informed* ASR-based approach that has access to the vocabulary of the test material (IASR).
- An *uninformed* ASR (UASR) that does not have access to the text vocabulary.

These are described in more detail in the following sections.

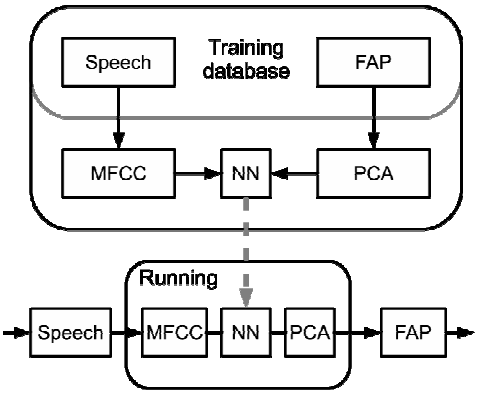


Figure 2. Structure of direct conversion

3.1 Direct conversion

The direct approach for audio-to-visual speech conversion was first presented in (Takács et al. 2006). A voice segment, from a different speaker used in training, was parameterized into MFCCs and a back-propagation neural network was used to estimate the best facial parameters to accompany this (novel) speech segment.

3.2 ASR-based conversion

For the ASR-based approaches, a Weighted Finite State Transducer — Hidden Markov-Model (WFST-HMM) decoder is used. Specifically, a system known as VOXerver (Mihajlik et al. 2007) is used, which can run in one of two modes: *informed*, which exploits knowledge of the vocabulary of the test data, and *uninformed*, which does not. Incoming speech is converted to MFCCs, after which blind channel equalization is used to reduce linear distortion in the cepstral domain (Mihajlik et al. 2005). Speaker independent cross-word decision-tree based triphone acoustic models previously trained using the MRBA Hungarian speech database (http://alpha.tmit.bme.hu/ speech/hdbMRBA.php) were applied.

The uninformed ASR system uses a phoneme-bigram phonotactic model to constrain the decoding process. The phoneme-bigram probabilities were estimated from the MRBA database. In the informed ASR system, a zerogram word language model was used with a vocabulary size of 120 words. Word pronunciations were determined automatically as described in the paper by Mihajlik, Révész and Tatai (2002).

In both types of speech recognition approaches, the WFST-HMM recognition network was constructed offline using the AT&T FSM toolkit (Mohri et al 2002). In the case of the informed system, phoneme labels were projected on to the output of the transducer instead of word labels. The precision of the segmentation was 10 ms.

3.3 Parameter generation

3.3.1 Viseme interpolation

To compare the direct and indirect audio-to-visual conversion systems, a standard approach for generating visual parameters is to first convert a phoneme to its equivalent viseme via a look up table, then linearly interpolate the viseme targets. This approach to synthesizing facial motion is naive in that coarticulation effects are ignored, but it does provide a baseline on expected performance (i.e., the worst-case scenario).

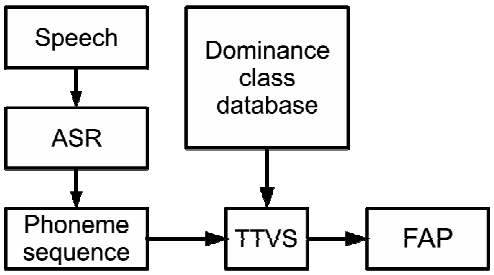


Figure 3. Modular ATVS consists of an ASR subsystem and a text to visual speech subsystem

3.3.2 Modular ATVS

To account for coarticulation effects, a more sophisticated interpolation scheme is required. In particular the relative dominance of neighboring speech segments on the

articulators must be taken into account. Speech segments can be classified as dominant, uncertain or mixed according to the level of influence exerted by the local neighborhood. To learn the dominance functions, an ellipsoid is fitted to the lips of a speaker in a video sequence articulating Hungarian triphones. To aid the fitting, the speakers wore a distinctly colored lipstick. Dominance functions are estimated by the variance of visual data in a given phonetic neighbourhood set. The learned dominance functions are used to interpolate between the visual targets derived from the ASR output. We used the implementation of László Czap and János Mátyás (2005) here which produced Poser script. FAPs are extracted from this format by the same workflow as from an original recording.

3.4 Rendering module

The visualization of the output of the ATVS methods is common to all approaches. The output from the ATVS modules are facial animation parameters (FAPs), which are applied to a common head model for all approaches. Note, although better facial descriptors than MPEG-4 are available, MPEG-4 is used here because our motion capture system does not provide more detail than this. The rendered video sequences are created from these FAP sequences using the Avisynth (http://avisynth.org) 3D face renderer. As the main components for the framework are common between the different approaches, any differences are due to variations in the AV mapping methods. Actual frames are shown on Figure 4.

4 Evaluation

Implementation specific noncritical behaviour (e.g., articulation amplitude) should be normalized to ensure that the comparison is between the essential qualities of the methods. To discover these differences, a preliminary test was completed.

4.1 Preliminary test

To tune the parameters of the systems, seven videos were generated by each of the five mapping methods, and some sequences were re-synthesized from the original facial motion data. All sequences started and ended with a closed mouth, and each contained between 2-4 words. The speaker used in all of the tests was not the same speaker used in training the audio-to-visual-mapping. The videos were presented in a randomized order to 34 viewers who were asked to rate the quality of the systems using an opinion score (1–5). The results are shown in Table 1.

Table 1. Results of preliminary tests used to tune the system parameters. Shown are the average and standard deviation of scores

Method	Average score	STD
UASR	3.82	0.33
Original	3.79	0.24
Linear	3.17	0.4
Direct	3.02	0.41
IASR	2.85	0.72

Figure 4. An example of the importance of correct timing. Frames of the word "Október" show timing differences between methods. Note that direct conversion received best score, even though it does not close the lips on bilabial but closes on velar, and it has problems with lip rounding. The row "Original" is the face of the actual talker, not the reference test material. The reference test material is articulated by the professional lip-speaker who speaks in the training database. This "Original" row is to show the actual timing of the visual part of the speech input.

The results (Table 1) were unexpected, IASR, which uses a more sophisticated coarticulation model was expected to be one of the best performing systems; however, it came in last. Closer examination of the scores revealed that the lower scores were related to poorer audiovisual synchrony of IASR than for UASR. A qualitative difference between the direct and indirect approaches was the degree of mouth opening — the direct approach tended to open the mouth on average 30% more than the indirect approaches. Consequently, to bring the systems into the same dynamic range, the mouth opening for the direct mapping was damped by 30%.

Subjective scores increased approximately 0.8 points, which revealed an interesting difference between the optimal articulation for intelligibility for hearing-impaired and naturalness for hearing people. The synchrony of the ASR-based approaches was checked for systemic errors (constant or linearly increasing delays) using cross correlation of locally time shifted windows, but no systematic patterns of errors were detected.

5 Results

5.1 ASR subsystem

The quality of the ASR-based approach was affected by the recognised phoneme string. This typically is 100% for the informed system as the test set consists only of a small number of words (months of the year, days of the week, and numbers under 100), while the uninformed system has a typical error rate of 25.21%. Despite this, the ATVS using this input performs surprisingly well. The likely reason might be the pattern of confusions – some of the phonemes that were confused acoustically appeared visually similar on the lips. The error expressed as the difference between the visemes (used in linear interpolation method) of the confused phonemes resulted in 45% of randomly chosen confusion differences.

A second factor that affects the performance of the ASR-based approaches is precision of segmentation. Generally the uninformed systems are more precise on the average than the informed systems. The precision of the segmentation can severely impact on the subjective opinion scores. Therefore we first attempted to quantify these likely sources of error.

The informed recognition system is similar in nature to forced alignment in standard ASR tasks. For each utterance, the recognizer is run in forced alignment mode for all of the vocabulary entries. The main difference between the informed and the uninformed recognition process is the different Markov state graphs for recognition. The informed system is a zerogram without loopback, while the uninformed graph is a bigram model graph, where the probabilities of the connections depend on language statistics.

While matching the extracted features with the Markovian states, differences were noted in both scenarios. However, the uninformed system allowed for different phonemes outside of the vocabulary to minimize the accumulated error. For the informed system only the most likely sequence was allowed, which can distort the segmentation — see Figure 4 for an example where the speaker mispronounces the word “Hatvanhárom” (hOtvOnha:rom, “63” in Hungarian). The (mis)segmentation of OtvO means the IASR AVTS system opened the mouth after the onset of the vowel. Human perception is very sensitive to this type of error and this severely impacted the perceived quality. Without forcing the vocabulary, a system may ignore one of the consonants, but open the mouth at the correct time. Note that generalising this phenomena is beyond the scope of this paper. We have demonstrated that this is a problem with certain implementations of HMM-based ASR. Alternative, more robust implementations might alleviate these problems.

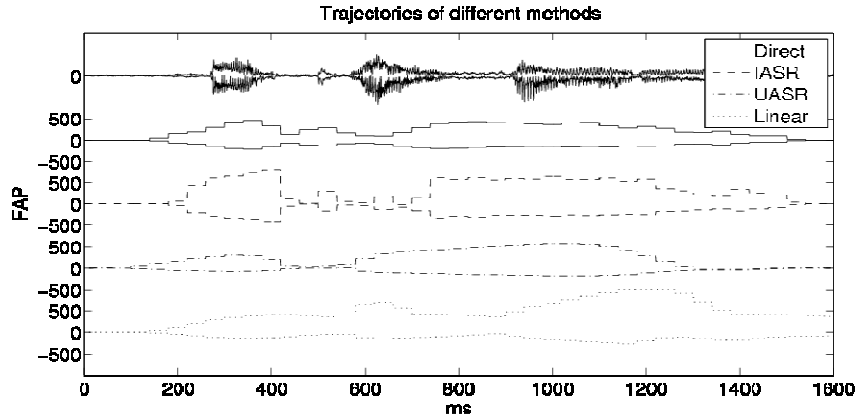


Figure 5. Trajectory plot of different methods for the word "Hatvanhárom" (hOtvOnha:rom). Jaw opening and lip opening width is shown. Note that the speaker did not pronounce the utterance perfectly, and the informed system attempts to force a match with the correctly recognized word. This leads to time alignment problems.

5.2 Subjective opinion scores

The test setup was similar to the previously described preliminary test used to tune the system. This time, there were 58 viewers, and there was no qualitative opinion survey given.

The results of the opinion score test is in Table 2. The advantage of direct conversion against UASR is approaching significance (p = 0.0512), as well as the difference between the original speech and the direct conversion (p = 0.06). However, UASR was significantly worse than the original speech (p = 0.00029). When compared to the preliminary test, these results also showed that with respect to naturalness, excessive articulation was not significant. The advantage of correct timing over correct phoneme string was also significant.

Table 2. Results of opinion scores, average and standard deviation

Method	Average score	STD
Original facial motion	3.73	1.01
Direct conversion	3.58	0.97
UASR	3.43	1.08
Linear interpolation	2.73	1.12
IASR	2.67	1.29

Note that the linear interpolation system is returning better quality ASR results, but still performs significantly worse than the average of other ASR-based approaches. This demonstrates the importance of correctly handling viseme dominance and viseme neighborhood sensitivity in ASR-based ATVS systems.

5.3 Intelligibility

Intelligibility was measured with a test of recognition of video sequences without sound (Table 3). This is not the popular modified rhyme test but for our purposes with hearing impaired viewers, it is deemed to be more relevant. The 58 test subjects had to guess which word was said from a given set of 5 other words. The sets were numbers, names of months and the days of the week. All words were said twice. The sets were presented as intervals to eliminate the memory test from the task. This task models the situation of being hearing-impaired or in a very noisy environment, where an ATVS system can be used. It is assumed that the context is known, so keyword spotting is the closest task to the problem.

Table 3. Results of recognition tests, average and standard deviation of success rate in percent. Random pick would give 20%.

Method	Precision	STD
IASR	61%	20%
UASR	57%	22%
Original motion	53%	18%
Cartoon	44%	11%
Direct conversion	36%	27%

The performance of the audio-to-visual speech conversion methods were reversed in this task compared to naturalness task. The main result here was the dominance of ASR-based approaches (see Table 2), and the insignificance of the difference between informed and uninformed ATVS results (p = 0.43), which may deserve further investigation. Note, when synchrony is not an issue without voice, the IASR is the best.

Öhman and Salvi published a comparison study of two ATVS implementations used in SYNFACE (Beskow et al. 2004) and internally at KTH (Öhman–Salvi 1999). The compared methods are the HMM rule-based system of SYNFACE and an approximation of this method learned by a neural network from audio speech. Manually set articulation parameters were used asa reference to achieve an ideal result. The results from their paper are in Table 4.

In comparison with Öhman and Salvi, where intelligibility was tested similarly, the manually tuned optimal rule based facial parameters were close to our IASR since there were no recognition errors, and without voice, the time alignment quality was not important, and our TTVS is rule-based. Their HMM test is similar to our UASR, because both are without vocabulary, both target a time aligned phoneme string to be converted to facial parameters, and our ASR is HMM-based. Their ANN system is very close to our direct conversion except for the training set, it is a standard speech database audio, and a rule-based calculated trajectory video data, while our system is trained on actual recordings from a professional lip-speaker.

Despite these differences, the results concerning intelligibility were close to each other (see Table 3). This is a validation of our results.

Table 4. Comparison to the results of Öhman & Salvi, a HMM and rule based systems intelligibility test. Intelligibility of corresponding methods are similar.

Our results		**Öhman & Salvi**	
Methods	**Prec.**	**Methods**	**Prec.**
IASR	61%	Ideal	64%
UASR	57%	HMM	54%
Direct	36%	ANN	34%

6 Conclusion

This paper has presented a comparative study of audio-to-visual speech conversion methods. We have presented a comparison of our direct conversion system with conceptually different conversion solutions. A subset of our results correlate with already published results, validating the approach.

We observed that the synchrony over phoneme precision in an ASR based ATVS system was a key component. There are publications demonstrating the importance of correct timing in different aspects (Czap & Mátyás 2005; Bailly et al. 2008; Feldhoffer et al. 2007), but our results explicitly showed that more accurate timing achieves much better subjective evaluation than a more accurate phoneme sequence. Also, we have shown that in the aspect of subjective naturalness evaluation, direct conversion (trained on professional lip-speaker articulation) is a method which produced the highest opinion score (95.9%) of an original facial motion recording with lower computational complexity than ASR-based solutions. For tasks where intelligibility is important (support for individuals who are hearing impaired, visual information in noisy environment) modular ATVS is the best approach among those presented. Our mission of aiding people who are hearing impaired requires us to consider using ASR-based components. For naturalness (animation, entertaining applications), direct conversion is a good choice. For both aspects, UASR gives relatively good, but not outstanding results.

Since the topic of ATVS comparison is very diversified, we know that our results are hard to reproduce. We have tried to facilitate this process by including an appendix of technical results. In addition we encourage visual speech research laboratories to test their systems with standardized output, which would make the comparison not just easier, but reproducible.

7 Technical details

Marker tracking was done for MPEG-4 FP 8.8 8.4 8.6 8.1 8.5 8.3 8.7 8.2 5.2 9.2 9.3 9.1 5.1 2.10 2.1. During synthesis, all FAPs connected these FPs were used except depth information: open_jaw, lower_t_midlip, raise_b_midlip, stretch_l_cornerlip, stretch_r_cornerlip, lower_t_lip_lm, lower_t_lip_rm, raise_b_lip_lm, raise_b_lip_rm,

raise_l_cornerlip, raise_r_cornerlip, lower_t_midlip_o, raise_b_midlip_o, stretch_l_cornerlip_o, stretch_r_cornerlip_o, lower_t_lip_lm_o, lower_t_lip_rm_o, raise_b_lip_lm_o, raise_b_lip_rm_o, raise_l_cornerlip_o, raise_r_cornerlip_o. Inner lip contour is estimated from outer markers.

Yellow paint was used to mark the FP locations on the face of the recorded lip-speaker. The video recording was 576i PAL (576 × 720 pixels, 25 frame/sec, 24 bit/pixel). The audio recording was mono 48 kHz 16 bit in a silent room. Further conversions depended upon the actual method.

Marker tracking was based on color matching and intensity localization frame and the location was identified by the region. In overlapping regions, the closest location on the previous frame was used to identify the marker. A frame with a neutral face was selected to use as the reference for FAPU measurement. The marker on the nose was used as reference to eliminate head motion.

Direct conversion used a modification of Davide Anguita's Matrix Backpropagation which enables real-time work also. The neural network used 11 frame long window on the input side (5 frames to the past and 5 frames to the future), and 4 principal component weights of FAP on the output. Each frame on the input was represented by a 16 band MFCC feature vector. The training set of the system contained stand-alone words and phonetically balanced sentences.

In the ASR, the speech signal was converted to a frequency of 16 kHz. MFCC (Mel Frequency Cepstral Coefficients)-based feature vectors were computed with delta and delta-delta components (39 dimensions in total). The recognition was performed on a batch of separated samples. Output annotations and the samples were joined, and the synchrony between labels and the signal was checked manually.

The visemes to the linear interpolation method were selected manually for each viseme in Hungarian from the training set of the direct conversion. Visemes and phonemes were assigned by a table. Each segment was a linear interpolation from the actual viseme to the next one. Linear interpolation was calculated in the FAP representation.

TTVS is a Visual Basic implemented system with a spreadsheet of timed phonetic data. This spreadsheet was changed to ASR output. Neighborhood dependent dominance properties were calculated and viseme ratios were extracted. Linear interpolation, restrictions concerning biological boundaries and median filtering were applied in this order. The output was a Poser data file which was applied to a model. The texture of the model was modified to black skin and differently colored MPEG-4 FP location markers. The animation was rendered in draft mode, with the field of view and resolution of the original recording. Marker tracking was performed, as described above, with the exception of the differently colored markers. FAPU values were measured in the rendered pixel space, and FAP values were calculated from FAPU and tracked marker positions.

This was done for both ASR runs, uninformed and informed.

The test material was manually segmented to 2-4 word units. The lengths of the units were around 3 seconds. The segmentation boundaries were listed and the video cut was automatically done with an Avisynth script. We used an MPEG-4 compatible head model renderer plugin for Avisynth, with the model “Alice” of XFace project. The viewpoint and the field of view was adjusted to have only the mouth on the screen in frontal view.

During the test, the subjects watched the videos fullscreen and used headphones.

8 Acknowledgment

The authors thank Péter Mihajlik for the support of the speech recognizer and the useful remarks, László Czap and János Mátyás for the TTVS, György Takács, Márton Péri, PPKE ITK and its students who helped at the tests.

References

Al Moubayed, S., De Smet, M. and Van Hamme, H. 2008. Lip synchronization: from phone lattice to pca eigen-projections using neural networks. In *Proceedings of Interspeech 2008*. Brisbane, Australia. 2016-2019.

Bailly, G., Govokhina, O., Breton, G. and Elisei, F. 2008. A trainable trajectory formation model td-hmm parameterized for the lips 2008 challenge. In *Proceedings of Interspeech 2008*. Brisbane, Australia. 2318-2321.

Beskow, J., Karlsson, I., Kewley, J. and Salvi, G. 2004. Synface - a talking head telephone for the hearing-impaired. *Computers Helping People with Special Needs*, 1178-1186.

Czap, L. and Mátyás, J. 2005. Virtual speaker. *Híradástechnika Selected Papers*, LX/6: 2-5.

Feldhoffer, G. 2007. Speaker independent continuous voice to facial animation on mobile platforms. In *49th International Symposium ELMAR*. Zadar, Croatia. 155-158.

Feldhoffer, G., Bárdi, T., Takács, Gy. and Tihanyi, A. 2007. Temporal asymmetry in relations of acoustic and visual features of speech. In *15th European Signal Processing Conf.* Poznan, Poland. 2341-2345.

Hofer, G., Yamagishi, J. and Shimodaira, H. 2008. Speech-driven lip motion generation with a trajectory hmm. In *Proceedings of Interspeech 2008*. Brisbane, Australia. 2314-2317.

Mihajlik, P., Fegyó, T., Németh, B. and Trón, V. 2007. Towards automatic transcription of large spoken archives in agglutinating languages: Hungarian asr for the MALACH project. In *Speech and Dialogue: 10th International Conference*. Pilsen, Czech Republic. 342-349.

Mihajlik, P., Révész, T. and Tatai, P. 2002. Phonetic transcription in automatic speech recognition. *ACTA LINGUISTICA HUNGARICA*, 49(3-4): 407–425.

Mihajlik, P., Tobler, Z., Tüske, Z. and Gordos, G. 2005. Evaluation and optimization of noise robust front-end technologies for the automatic recognition of hungarian telephone speech. In *Interspeech 2005 - Eurospeech: 9th European Conference on Speech Communication and Technology*. Lisboa, Portugal. 2677-2680.

Mohri, M., Pereira, F. C. N. and Riley, M. 2002. Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, 16: 69-88.

Öhman, T. and Salvi, G. 1999. Using hmms and anns for mapping acoustic to visual speech. *TMH-QPSR*, 40: 45-50.

Takács, Gy., Tihanyi, A., Bárdi, T., Feldhoffer, G. and Srancsik, B. 2006. Speech to facial animation conversion for deaf customers. In *4th European Signal Processing Conf.* Florence, Italy.

Theobald, B., Fagel, S., Elsei, F. and Bailly, G. 2008. LIPS2008: Visual speech synthesis challenge. In *Proceedings of Interspeech 2008.* Brisbane, Australia. 1875-1878.

IMPROVING NATURALNESS OF VISUAL SPEECH SYNTHESIS

László Czap* and János Mátyás**
***University of Miskolc, Department of Automation, Hungary**
****North Hungarian Training Centre, Miskolc, Hungary**
e-mail: czap@mazsola.iit.uni-miskolc.hu, matyasj@lab.hu

Abstract

Facial animation has progressed significantly over the past few years and a variety of algorithms and techniques now make it possible to create highly realistic characters. Based on the author's speechreading study and the development of 3D modelling, a Hungarian talking head has been created. Our general approach is to use both static and dynamic observations of natural speech to guide the facial modelling. Evaluation of Hungarian consonants and vowels served the classification of visemes - the smallest perceptible visual units of the articulation process. A three level dominance model has been introduced that takes coarticulation into account. Each articulatory feature has been grouped to *dominant*, *flexible* or *uncertain* classes. Analysis of the standard deviation and the trajectory of the features assisted the evaluation process. The acoustic speech and the articulation are linked with each other by a synchronising process. There are a couple of features added to improve the naturalness of articulation:

1. Pre-articulation. Prior to utterance a silence period is inserted – imitating breathing by opening the mouth – then the first dominant viseme is progressed from the neutral starting position.
2. A filtering and smoothing algorithm has been developed for adaptation to the tempo of either the synthesized or natural speech.
3. Eye blink, gaze, head movement, and eyebrow rising can be controlled semi-randomly or through special commands.
4. Basic emotions defined by Ekman can be expressed in a scalable manner.

1 Introduction

The intelligibility of speech can be improved by showing the articulation of the speaker. This visual support is essential in a noisy environment and for people with hearing impairment. An artificial talking head can be a natural supplement to sophisticated acoustic speech synthesis. The pioneering work of face animation for modelling articulation started decades ago. The development of 3D body modelling, the evolution of computers and advances in the analysis of human utterances has enabled the development of realistic models. Teaching hearing impaired people to speak can be aided by an accurately articulating virtual speaker, which can make its

face transparent and can show details of the movements associated with pronunciation (Figure 1).

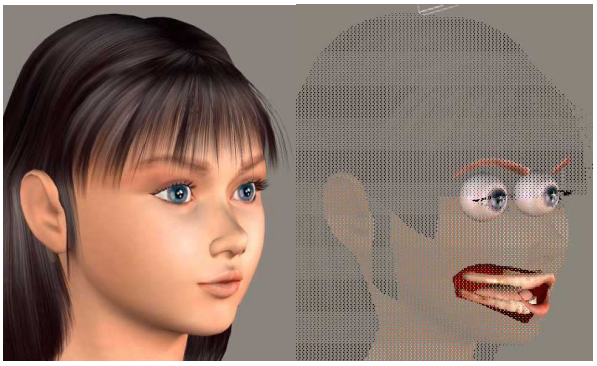


Figure 1. Photorealistic and transparent visualization

Since the last decade, the area of visual speech has been developing dynamically, with more and more applications being developed. Existing systems are focusing on high quality modelling of articulation, but have been restricted by number of polygons issues, e.g. simulation of hair falling needs more computation effort than the articulation itself.

A visual feature database for speech-reading and for the development of 3D modelling has been developed and a Hungarian talking head has already been created (Czap 2004). In this research, the general approach was to use both static and dynamic observations of natural speech to guide facial animation. A three-level dominance model has been introduced that takes co-articulation into consideration. Each articulatory feature has been grouped into one of three classes: dominant, flexible or uncertain. Analysis of the standard deviation and the trajectory of features guided the evaluation process. The acoustic features of speech and the articulation are then linked to each other by a synchronising process.

2 Research aim

This research is designed to model photorealistic appearance and sophisticated human-like articulation with close to natural movements of the speaker triggering four additional research results:

1. Pre-articulation. Prior to an utterance, a silent period is inserted – imitating breathing by opening the mouth – then the first dominant viseme is moved from the neutral starting position.

2. Realizing the temporal asynchrony effect. A filtering and smoothing algorithm has been developed for adaptation to the tempo of either the synthesized or natural speech.

3. Head movement, gaze, eyebrow rising, and eye blink. An algorithm was developed to semi-randomly and manually control the former movements.

4. Emotion movements. Following the definitions of Ekman, a scalable and blended method was developed to express emotions.

3 Method and material

The first visual speech synthesizers were based on a 2D head model, accessing previously stored images of a speaker. Phasing between these frames was sometimes produced by image morphing (Cosatto et al. 1998). However, a 2D model is not sufficient for providing head movements, gestures and emotions, but 3D models can simulate facial expressions by tensing muscles. A 3D model produces realistic results, but the replication of real muscular tensions by a person is difficult. Surface models using textured polygons seem to be promising . Their features can be analysed by assessing human speakers (Massaro 1998, Bernstein et al. 1996).

In this line of research, a 3D transformation of a geometric surface model was used. The deformation based articulation was translated into a parametric model to overcome the restrictions of the morphing technique. Facial movements are not carried out by deformations of the face, instead a collection of polygons is manipulated using a set of parameters. This process permits control of a wide range of motions using a set of features associated with different articulation functions. These parameters can be directly matched to particular movements of the lip, tongue, chin, eyes, eyelids, eyebrows and the whole face.

The visual representation of the speech sound (mostly representing the phoneme) is called a *viseme*. A set of visemes has fewer elements than that of phonemes. The static positions of the speech organ for the production of Hungarian phonemes can be found in seminal works. Figure 2 shows the similarity of visemes using a speaker's photograph (Bolla 1995) compared to those produced by a 3D model.

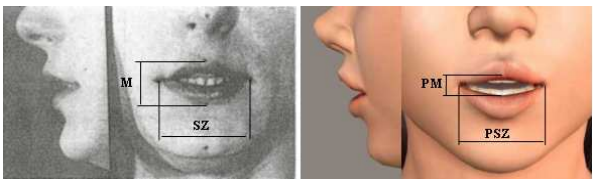


Figure 2. A photograph of a speaker and the 3D model of the same viseme

The features of Hungarian visemes have been created using the word models of Molnár (1986). The main features of visemes have been adopted from the published sound maps and albums (Bolla 1980, 1995). These features were transformed using the parameters of the articulation model by Mátyás (2003). Features controlling the lips and tongue are crucial. Basic lip properties include the opening and width, their movement is related to lip rounding. The lip opening and the visibility of teeth are dependent upon jaw movement. The tongue is described by its horizontal and vertical position, its bending and the shape of the tongue tip (Figure 3).

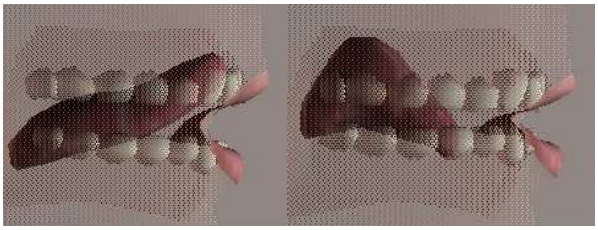


Figure 3. Illustrative tongue positions for sounds [n] (left) and [k, g] (right).

Based upon the static features, the articulation parameters characteristic to the stationary section of the viseme can be set.

4 Modelling dynamic operation

The dynamic features of conversational Hungarian have not been described yet. The current research begins to provide this description. The usefulness of motion phases represented in presentations containing photos on artculation of speech sounds (albums) are limited, and can be related only to the particular word given in the album. Dynamic analysis are taken from the authors' own studies in speechreading (Czap 2004). Specifically, this work provided the trajectories for the width and height of the oral cavity and the visibility of teeth and tongue. These data form the basis for movement between visemes.

Some features take on their characteristic value, while others do not reach their target value during the pronunciation. All features of the visemes (e.g., lip shape, tongue position) have been classified according to their dominance, with every articulation feature being assigned to a dominance class. This is different from the general approach where the visemes are grouped by their dominance only. For instance, the Hungarian visemes of the sounds [ɟ, c, j, ɲ] are dominant on lip opening and tongue position and are uncertain with respect to lip width. This categorization is based on the ranges provided by the speechreading data. The features of the parametric model can be divided into three grades:

- *dominant* – coarticulation has (almost) no effect on them,
- *flexible* – the neighbouring visemes affect them,
- *uncertain* – the neighbourhood determines the feature.

In addition to the range, the distribution of transitional and stationary periods of visible features helps to determine the grade of dominance.

The trajectory of viseme features can also be essential for determining dominance classes. Figure 4a shows the trajectory of inner lip width (horizontal axis) and lip opening (vertical axis) of the viseme [e:]. These curves cannot be traced one by one but they go through a dense area regardless of the starting and final states. The dominant nature of the vowels' lip shape is obvious. In contrast, uncertain features do not provide a consistent pattern. The trajectory of [h] can be seen in Figure 4b. (To be able to track them, only a few curves are represented.)

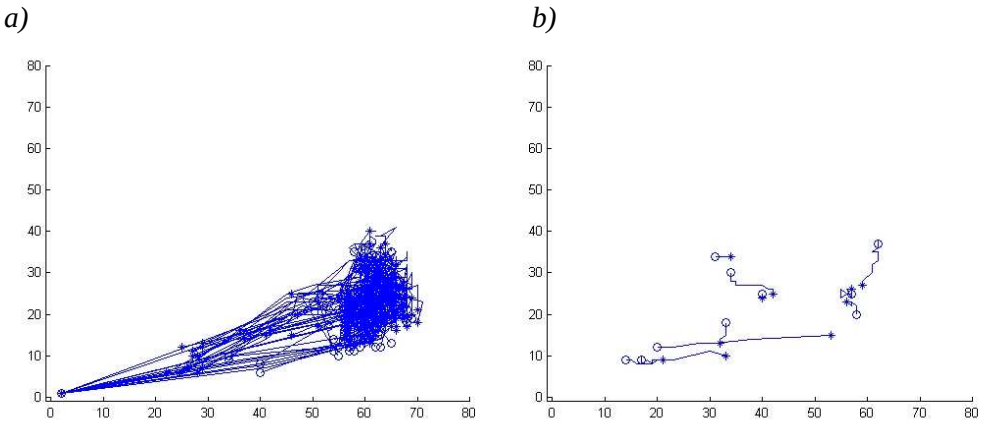


Figure 4. Trajectory of lip sizes in pixels of viseme [**e:**] (left) and [**h**] (right)

The dominance grade of the previously arranged features, which considers the deformability and context dependence of the viseme, controls the interpolation of features. Forward and backward co-articulation rules are applied in a way that articulation parameters are influenced by less elastic properties.

5 Pre-articulation

Prior to an utterance, there is an approximately 300 ms silence period is inserted – designed to imitate breathing through the mouth – then the first dominant viseme is moved from the neutral starting position. Because of this pre-articulatory movement, the sound is produced as in natural speech. If the last sound of the sentence was bilabial, then the mouth would be slightly opened after the sound fade (post-articulation).

6 Adapting to the tempo of speech and filtering

During the synchronization to human or synthesized speech, we have encountered different speech tempos. When speech is slow, viseme features approach their nominal value, while fast speech is articulated with less precision in natural speech. For flexible features, the round off is stronger in fast speech. A median filter is applied for interpolation of flexible features: the values of neighbouring frames are sorted and the median is chosen. A feature is formed by the following steps:

• linear interpolation among values of dominant and flexible features, neglecting uncertain ones,

• median filtering is performed when flexible features are juxtaposed,

• values are then filtered by the weighted sum of the two previous frames, the actual and the next one.

The weights of the filter are fixed, so knowledge of speech tempo is not needed. The smoothing filter refines the movements and reduces the peaks during fast speech. By considering the two previous frames, the timing asymmetry of articulation is approximated. Feldhoffer et al. (2007) have shown that the mouth

starts to form a viseme before the corresponding phoneme is audible. Filtering takes this phenomenon into account. Other improvements – as inserting a permanent phase into long vowels and synchronising phases of a viseme to a phoneme at several points – refine the articulation.

Figure 5. depicts the effect of median filtering and smoothing. In this example, the slow speech has twice as many frames as the fast one. The horizontal axis shows the number of frames, while vertical values represent the amplitude of the feature.

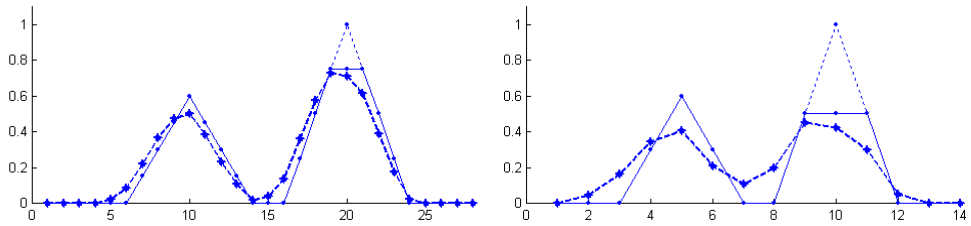


Figure 5. The interpolation of dominant (first peak) and flexible (second peak) features for slow (left) and fast (right) speech after linear interpolation (…), median filtering (___) and smoothing (---)

7 Improving naturalness

Visible features (e.g., nodding, eyebrow rising, blinking) are meaningful gestures, especially for people who are hard of hearing. In dialogues, gestures can support turn taking. For instance, the lift of eyebrows can indicate paying attention, while nodding can show acknowledgement. An algorithm for head movement and mimicry cannot easily be created using prosody, because the communicative context of the utterance needs to be taken into account. To link the automatically generated movements with the intent of the utterance, features can be managed manually using a graphical editor (e.g., lifting eyebrows or nodding at sentence accent, or controlling the eyes to imitate a glance into a paper).

By studying the head movements of professional speakers, moderate nodding, tilting and blinking were introduced. We have studied more than 15 minutes of speech from different announcers delivering the news on television. For each video frame, the eyes can be easily identified and traced using a conventional method for obtaining motion vectors for video compression, like MPEG standards. Horizontal movement reveals panning; while vertical movement means tilting. Unbalanced movement of left and right eyes indicates head inclination. When the difference of the absolute value of the region of interest on succeeding image frames is greater than a definite threshold, blinking is noted. These statistics form the basis for probability modelling of head movements. A series of subjective tests served to fine tune these parameters.

8 Facial gestures

Head movement, gaze, eyebrow rising, and eye blink can be controlled semi-randomly or manually. Automatic generation of facial gestures is organized in a semi-random manner. A rigid rule-based system would result in mechanical, boring and unnatural movements. Tilting and-nodding head movements are related to a short time (200 ms) average energy of the acoustic signal. For sentence accents, a downward head movement is observed. The bigger the average sound energy is, the higher the probability there is of downward head tilting. Moderate and slow head turning (or pan) and side inclination are controlled randomly. The amplitude of these head movements is not more than 2–3^{o}. Gaze is controlled by monitoring head movements so as to keep the face looking into the camera (i.e., the observer's eyes). Vertical and horizontal head movements are produced by moving both eyes in the opposite direction. The following equations show the control rule of vertical and horizontal direction of the left eye respectively:

$$D_{leyeh} = -D_{pan} \qquad D_{leyev} = -D_{tilt}$$

(The horizontal position of left eye is the opposite of the direction of pan head movement. The vertical position of left eye is compensating the tilt head movement.) Head movement is related to emotions as well, e.g., when expressing sadness, the head moves downward.

Based on observation of professional announcers and the existing literature, eyebrow movement is controlled by taking the probable strength relationship between the potential prosodic and visual cues into account. The interaction between acoustic intonation (F0) gestures and eyebrow movements has been studied in production, as in Cavé et al. (1996), for example. A preliminary hypothesis is that a direct coupling is very unnatural, but that prominence and eyebrow movement may co-occur. In our experiment, the brows were noted to rise subtly at the beginning of declarative sentences and then it approaches the neutral position. A larger raising movement is likely to be interpreted as surprise. Inner and outer eyebrows are to be risen independently when emotions are expressed. Eyebrows can be used to express anger, disgust and sadness (Figure 6).

Figure 6. Eyebrows of displaying surprise and worry

According to our observations, blinking occurred about every 1.5 to 3 seconds. Blinking is controlled semi-randomly. Higher average energy makes blinking more frequent. The probability of blinking is increased when long vowels or the first

vowel of the word is produced, as in Hungarian word-stress, which is always put on the first syllable of the word.

9 Expressing emotions

An entire industry has grown up around the study of facial expressions. Much of the research has followed the lead of Ekman (2001) and focuses on Darwin's first argument that facial expressions have evolved. When expressions are evolved, then humans must share a common, universal set. Many facial expressions are involuntary in animals, and probably in humans as well. As a consequence, much of the research has turned toward understanding what facial expressions show about the emotional state of the speaker. Emotional content should be obvious to others because everyone shares the same universal assortment of expressions. In multimodal speech, we can confirm or disprove the verbal message by using gestures and body language.

Researchers have found that subjects label certain expressions of emotion in the same way regardless of culture. This research has uncovered possibly six distinct emotions that are read with ease across cultures. These emotions include anger, disgust, happiness, sadness, fear, and surprise. Facial expressions signal a particular emotional state held by the speaker. This perspective leads to the conclusion that there are identifiable states that are easily perceived.

After Ekman (2001), in the presented system, the above mentioned basic emotions can be selected in a scalable manner (e.g., surprise causes inner and outer eyebrows raised, mouth open, eyes open, lips protruded, chin down and head upward). Figure 7 depicts four examples.

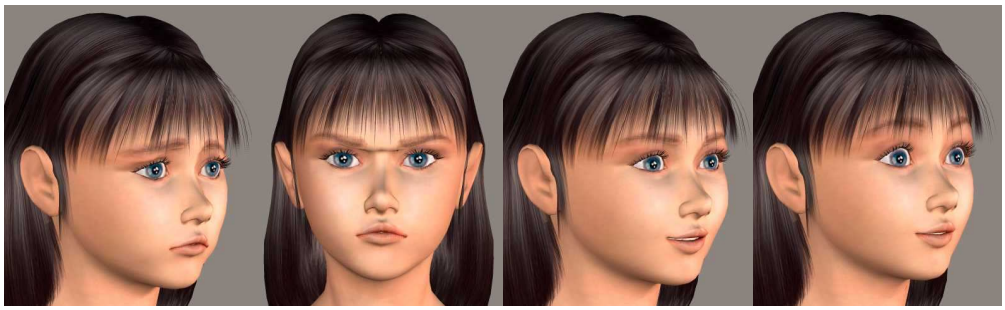


Figure 7. Expression of sadness, disgust, happiness and (nice) surprise

During the utterance of a sentence, facial expressions are progressing from a neutral look to the target display of emotion. Emotion can be controlled in a scalable and blended manner (Busso et al. 2008). E.g. 20&+30$ means 20% fear and 30% surprise, while 20*+30$ evolves 20% happiness and 30% surprise.

10 Conclusions

This paper describes the latest results of a larger research project that – in general – aims to create a Hungarian audio-visual text-to-speech system. Fine tuning of the probabilities for natural animation and avoiding mechanical, rule based repetition of gestures resulted from careful study of speech production. Pre- and post-articulation, median filtering for adaptation to the speech tempo, and filtering for temporal asymmetry of speech production have been introduced as directions for continued refinement of human-like articulation. In this phase, further manipulation of co-articulation is performed. Improving the naturalness and expressing the emotions will make the simulated performance more attractive.

Sample videos can be found: http://mazsola.iit.uni-miskolc.hu/~czap/mintak

References

Bernstein, L. E. and Auer, E. T. 1996. Word recognition in speechreading. In Stork, D. G. and Hennecke, M. E. (eds.): *Speechreading by humans and machines: models, systems and applications.* Berlin–Heidelberg: Springer-Verlag. 17-26.

Bolla, K. 1995. *A phonetic conspectus of Hungarian*. Budapest: Tankönyvkiadó.

Busso, C., Deng, Z., Neumann, U. and Narayanan, S. 2008. Learning expressive human-like head motion sequences from speech. In Deng, Z. and Neumann, U. (eds.): *Data-driven 3D facial animation.* London: Springer-Verlag.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F. and Espesser, R. 1996. About the relationship between eyebrow movements and F0 variations. In Bunnell, H. T. and Idsardi, W. (eds.): *Proceedings ICSLP 96.* Philadelphia, PA, USA. 2175-2178.

Cosatto, E. and Graf, H. P. 1998. Sample-based synthesis of photo-realistic talking heads. *Computer Animation,* 98: 103-110.

Czap, L. 2004. *Audiovizuális beszédfelismerés és beszédszintézis (= Audio-visual speech recognition and synthesis).* PhD thesis. Budapest University of Technology and Economics.

Ekman, P. 2001. Facial expressions. In Blakemore, C. and Jennett, S. (eds.): *Oxford companion to the body.* London: Oxford University Press.

Feldhoffer, G., Bárdi, T., Takács, Gy. and Tihanyi, A. 2007. Temporal asymmetry in relations of acoustic and visual features of speech. In *Proc. 15th European Signal Processing Conference.* Poznan, Poland. 2341-2345.

Massaro, D. W. 1998. *Perceiving talking faces*. Cambridge, Massachusetts–London, England: The MIT Press. 359-390.

Mátyás, J. 2003. *Vizuális beszédszintézis (= Visual speech synthesis)*. MSc thesis. University of Miskolc.

Molnár, J. 1986. *A magyar beszédhangok atlasza (= The map of Hungarian speech sounds).* Budapest: Tankönyvkiadó.

EXPLORING DIFFERENCES BETWEEN PHONETIC CLASSES IN SLEEP APNOEA SYNDROME PATIENTS USING AUTOMATIC SPEECH PROCESSING TECHNIQUES

Jose Luis Blanco, Rubén Fernández,
Eduardo López and Luis Alfonso Hernández
Departamento de Señales, Sistemas y Radiocomunicaciones,
Universidad Politécnica de Madrid, Spain
e-mail: jlblanco@gaps.ssr.upm.es, ruben@gaps.ssr.upm.es,
eduardo@gaps.ssr.upm.es, luis@gaps.ssr.upm.es

Abstract

Early detection of obstructive sleep apnoea syndrome (OSA syndrome) using automatic speech processing techniques has become of great interest because the current diagnostic methods are expensive and time-consuming. Pioneering research in this field has recently yielded some promising results based on differences noted when comparing voices recorded from patients suffering from apnoea and those from healthy people. However, the relationship between this condition and the noted vocal abnormalities is still unclear, because the speech signals have not been systematically described. Most of the information used to describe the vocal effects of apnoea comes from a perceptual study where phoneticians were asked to compare voices from apnoea patients with a control group. These results revealed abnormalities in articulation, phonation and resonance.

This work is part of an on-going collaborative project between the medical and signal processing communities to promote new research efforts on automatic OSA diagnosis. In this paper, we explore the differences noted in phonetic classes (inter-phoneme) across groups (control/apnoea) and analyze their utility for OSA detection. Using statistical models, inter-phoneme scores were evaluated to quantify the predictive capability associated with each phonetic class for identifying this pathology. A global predictive power of 72% was obtained by combining inter-phoneme scores from four phonetic classes. We also compared these scores to identify the most discriminative phonetic classes. This process will help us improve our overall understanding of the effects of OSA on speech. Finally, using the Kullback-Leibler distance, significant differences were found for vowel production in nasal vs. non-nasal contexts. This was probably the result of the abnormal coupling of the oral and nasal cavities observed in apnoea patients. This finding represents a relevant result for future research.

1 Introduction

Obstructive sleep apnoea (OSA) is a highly prevalent disease (Fox & Monoson 1989), and it is estimated that in middle-age adults many as 9% of women and 24% of men are affected, undiagnosed and untreated (Lee et al. 2008). This disorder is characterized by recurring episodes of sleep-related collapse of the upper airway at the level of the pharynx and it is usually associated with loud snoring and increased daytime sleepiness. OSA is a serious threat to an individual's health if not treated. This condition is a risk factor for hypertension and, possibly, cardiovascular diseases (Coccagna et al. 2006). It is usually a factor in traffic accidents caused by somnolent drivers (Lee et al. 2008; Coccagna et al. 2006; Lloberes et al. 2000), and it can lead to a poor quality of life and impaired work performance. Current diagnostic procedures require a full overnight sleep study to confirm the presence of the disorder. This procedure involves recording neuroelectrophisiological and cardiorespiratory variables (ECG), which then results in a 90% accuracy rate in detecting OSA. Nevertheless, this is an expensive and time-consuming diagnostic protocol, and, in some countries such as Spain, patients have to remain on a waiting list for several years before the test can be completed. This is because the demand for consultations and diagnostic studies for OSA has significantly increased (Lee et al. 2008). There is, therefore, a strong need for methods of early diagnosis of apnoea patients in order to alleviate these considerable delays and inconveniences.

In over 25 years of research, a number of factors have been related to the upper airway (UA) collapse during sleep-time. Essentially, pharyngeal collapse occurs when the normal reduction in pharyngeal dilator muscle tone at the onset of sleep is superimposed on a narrowed and/or highly compliant pharynx. This suggests that OSA may be a heterogeneous disorder rather than a single disease, involving the interaction of anatomic and neural state-related factors resulting in pharyngeal collapse. However, it is interesting to consider that OSA is an anatomic illness that might have been favoured by evolutionary adaptations in the human's upper respiratory tract (Davidson 2003). Anatomic changes include shortening of the maxillary, ethmoid, palatal and mandibular bones; acute oral cavity-skull base angulation, pharyngeal collapse with anterior migration of the foramen magnum, posterior migration of the tongue into the pharynx with descent of the larynx, and shortening of the soft palate with loss of the epiglottic–soft palate lock-up.

Phoneticians have also taken a look into OSA from their own perspective (for instance Fox & Monoson 1989) and concluded that although articulatory, physiologic and acoustic anomalies are somewhat unclear, results involving combinations of factors have some explanatory power. Nevertheless, such an anomaly should result not only in respiratory, but also in speech dysfunction. Consequently, the occurrence of speech disorder in the OSA population should be expected, and it would likely involve anomalies in articulation, phonation and resonance. The most representative of these abnormalities are described in Section 2.

In this paper, we investigate the acoustic characteristics of speech in patients suffering from OSA by using techniques taken from automatic speech and speaker

recognition. Using generative statistical models to describe the acoustic space, we explore the differences between phonetic classes and their possible application to automatic detection of OSA. These phonetic classes have offered a good trade-off between data complexity and recognition rate in speaker verification scenarios (Hébert & Heck 2003), especially when sparse data are available. The differences in the variability observed between and within a group of healthy speakers and those suffering from OSA are significant enough to motivate further research and reflect what phoneticians had observed in their previous experiments.

The remainder of this paper is organized as follows: in Section 2 the physiological and acoustic characteristics described in the previous literature on speakers suffering from severe apnoea syndrome are reviewed. On the basis of the limited information available about the side-effects of this condition on speech, a specific speech corpus was designed to test differences between both a normal and patient population. The design of this corpus, i.e., a brief analysis of the sentences it contains, is presented in Section 3; while Section 4 briefly describes the characteristics of the recorded speech database and provides several physical characteristics of the speakers in both groups. In Section 5, our approach, based on modelling the acoustic space using statistical models is presented. Once the experimental framework has been set, Section 6 describes the actual phonetic classes identified and provides details on their representation using statistical models. In Section 7, experimental results exploring differences between OSA and healthy speakers are presented using inter-phoneme and intra-phoneme scores. Finally, some conclusions and a brief outline on the future work are provided in Section 8.

2 Physiological and acoustic characteristics of OSA speakers

Currently, the articulatory/physiological settings as well as the acoustic characteristics of speech in speakers suffering from apnoea syndrome (for simplicity we will refer them as apnoea speakers), are still unclear. Most of the more valuable information in this field can be found in Fox and Monoson's work (1989), where a perceptual study with skilled judges was presented comparing voices from apnoea patients and a control group (hereafter referred to as "healthy" speakers). This study revealed differences between both groups of speakers, however acoustic cues for these differences were somewhat contradictory and unclear. What did seem to be clear was that speakers in the apnoea group exhibited abnormal resonances that might appear due to the altered structure or function of the upper airway. Theoretically this anomaly should result not only in a respiratory but also in a speech dysfunction, which is our primary hypothesis. The abnormalities previously identified are the following:

Articulatory anomalies: Fox and Monoson (1989) pointed out that neuromotor dysfunctions could be found in a sleep apnoea population as a "lack of regulated innervations to the breathing musculature or upper airway muscle hypotonus". This type of dysfunction is normally related to speech disorders, especially dysarthria. There are several types of dysarthria, each incorporating different acoustic features.

However, all types of dysarthria affect the articulation of consonants and vowels causing the slurring of speech. Another common pair of features in apnoea patients is hyper- and hypo-nasality, as well as a number of problems related to respiration.

Phonation anomalies: Phonation anomalies may appear due to the fact that heavy snoring in sleep apnoea patients can cause inflammation in the upper respiratory system and affect the vocal cords.

Resonance anomalies: The analysis of resonance characteristics for the sleep apnoea group in Fox and Monoson's work (1989) did not yield a clear conclusion. It was only recently that resonance disorders affecting speech quality have been associated with vocal tract damping features, distinct from airflow imbalance between the oral and nasal cavities. The term applied to this particular speech disorder is "cul-de-sac" resonance, and refers to a specific type of hyponasality. However, researches could only conclude that resonance abnormalities in apnoea patients could be perceived both as hyponasality (no nasalization is produced when the sound should be nasal) or hypernasality (nasalization is observed during production of non-nasal –voiced oral– sounds). Furthermore, and perhaps more importantly, speakers with apnoea seemed to exhibit smaller intra-speaker differences between non-nasal and nasal vowels due to this dysfunction, when vowels ordinarily require either a nasal or a non-nasal quality. Additionally, due to pharyngeal anomalies, differences in formant values can be expected. This was confirmed by Robb's work (Robb et al. 1997), in which vocal tract acoustic resonance was evaluated in a group of OSA males. Statistically significant differences were found in formant frequencies and bandwidth values between apnoea and healthy groups. In particular, the results of the formant frequency analysis showed that F1 and F2 values among the OSA group were generally lower than for the non-OSA group.

Finally, these anomalies can occur either in isolation or in combination. However, none of them was found to be sufficient on its own to allow accurate assessment of the OSA condition. In fact, all three descriptors were necessary to differentiate and predict whether the subject belonged either to the healthy or the OSA groups.

3 Speech corpus

The speech corpus was specifically designed to test differences between healthy people and those suffering from OSA. It contains four sentences in Spanish that are repeated three times by each speaker (Fernández et al. 2008). Keeping Fox and Monoson's work in mind, the sentences were designed so that they include instances of the following specific phonetic contexts:

• In relation to **articulatory anomalies** we collected voiced sounds affected by preceding phonemes that have their primary locus of articulation near the back of the oral cavity, specifically, velar phonemes, such as the Spanish velar approximant /g/. This anatomical region has been known to display physical anomalies in speakers suffering from apnoea (Davidson 2003). Thus, it is reasonable to suspect that different coarticulatory effects may occur with these phonemes in speakers with

and without apnoea. In particular, in our corpus, we collected instances of transitions from the Spanish voiced velar plosive /g/ to vowels, in order to analyse the specific impact of articulatory dysfunctions in the pharyngeal region.

• With regard to **phonation anomalies**, we included continuous use of voiced sounds to measure possible irregular phonation patterns related to muscular fatigue noted in apnoea patients.

• Finally, to look at **resonance anomalies**, we designed sentences that allowed intra-speaker variation measurements; that is, measuring differential voice features for each speaker, for instance to compare the degree of vowel nasalization within and without nasal contexts.

Moreover, all sentences were designed to exhibit a similar melodic structure, and speakers were asked to try reading them with a specific rhythm under the supervision of an expert. We followed this controlled rhythmic recording procedure hoping to minimise non-relevant inter-speaker linguistic variability. The sentences chosen were the following, with the different melodic groups underlined separately:

(1) Francia, Suiza y Hungría ya hicieron causa común.
'fraNθja 'sujθa i uŋ 'gri a ya i 'θje roŋ 'kaw sa ko 'mun

(2) Julián no vio la manga roja que ellos buscan, en ningún almacén.
xu 'ljan no 'βjo la 'maŋ ga 'ro xa ke 'e λoz 'βus kan en niŋ 'gun al ma 'θen

(3) Juan no puso la taza rota que tanto le gusta en el aljibe.
xwan no 'pu so la 'ta θa 'řo ta ke 'taN to le 'ɣus ta en el al 'xi βe

(4) Miguel y Manu llamarán entre ocho y nueve y media.
mi 'ɣel i 'ma nu λa ma 'ran 'eN tre 'o tʃo i 'nwe βe i 'me ðja

The first phrase was taken from the Albayzin database, a standard phonetically balanced database for Spanish (Moreno et al. 1993). It was selected because it contains an interesting sequence of successive /a/ and /i/ vowel sounds.

The second and third phrases, both negative, have a similar grammatical and intonation structure. They are potentially useful for contrastive studies of vowels in different linguistic contexts. Some examples of these contrastive pairs arise from comparing a nasal context, "m**a**ng**a** roj**a**" (*'maŋ ga 'řo xa*), with a neutral context, "t**a**z**a** rot**a**" (*'ta θa 'řo ta*). These contrastive analyses could be very helpful to confirm whether the voices of speakers with apnoea had an altered overall nasal quality and displayed smaller intra-speaker differences between non-nasal and nasal vowels due to velopharyngeal dysfunction.

The fourth phrase has a single and relatively long melodic group, containing largely voiced sounds. The rationale for this fourth sentence was that apnoea speakers usually show fatigue in the upper airway muscles. Therefore, this sentence might help us discover anomalies during the generation of voiced sounds. This

sentence also contains several vowel sounds embedded in nasal contexts that could be useful to study phonation and articulation of nasalized vowels. Finally, with regard to the resonance anomalies found in the literature and previously described, one of the possible traits of apnoea speakers is **dysarthria.** This last sentence could also be used to analyse dysarthric voices that typically show differences in vowel space when compared to healthy (control) speakers (Turner et al. 1995).

4 OSA database collection

The database, which in the rest of the paper will be referred to as OSA database, was recorded in the Respiratory Department at Hospital Clínico Universitario de Málaga, Spain. It contains the readings of 80 male subjects; half of them suffering from severe sleep apnoea (high Apnoea – Hipoapnoea Index values, AHI > 30), and the other half were either healthy subjects or had mild OSA (AHI < 10). Subjects in both groups had similar physical characteristics, such as age and Body Mass Index (BMI, i.e. weight divided by the square of height) - see Table 1.

Table 1. Distribution of healthy and pathological speakers in the OSA database

	Number	**Mean Age**	**Std. dev. Age**	**Mean BMI**	**Std. dev. BMI**
Control	40	42.2	8.8	26.2	3.9
Apnoea	40	49.5	10.8	32.8	5.4

Our selection of speakers for each group attempted to avoid the influence of the external predisposing factors associated with the condition. Such an approach ensures that the results are most likely related to group factors and can be generalized to a homogeneous population.

Moreover, speech was recorded using a sampling rate of 16 kHz in an acoustically isolated booth. The recording equipment consisted of a standard laptop computer with a conventional sound card equipped with a SP500 Plantronics headset microphone with A/D conversion and digital data exchange accomplished through a USB-port.

5 Statistical modelling of the acoustic space

The discrimination of normal and pathological voices using automatic acoustic analysis and speech recognition technology is becoming an alternative method of diagnosis for researchers in laryngological and speech pathologies, because of its nonintrusive nature and its potential for providing quantitative data relatively quickly. State-of-the-art speech recognition technology can be briefly described as the use of machine learning techniques to train a statistical model from acoustic features representing a known acoustic space (see [Huang et al. 2001] for a complete introduction to speech technology). These acoustic features are extracted from a training speech database where the speech from specific speakers is recorded and properly annotated. So, in **speaker recognition,** these acoustic features come from a

known speaker's voice, while in **speech recognition,** the acoustic space is generally covered by a set of phoneme-like units representing a given language. After training, the acoustic features coming from an unknown speaker or spoken sentence are recognized based on the likelihood scores obtained supposing that the unknown acoustic features were generated by a statistical model representing a particular speaker (speaker recognition) or linguistic unit (speech recognition). So, for example, in speaker recognition, a certain speaker is recognized when values from the acoustic features being tested are more likely (i.e. higher likelihood score) for the speaker's own statistical model, rather than any other model in the system.

Given this brief overview of speech recognition and the expected speech abnormalities in patients with apnoea syndrome, it can be seen that the use of this technology to explore differences between apnoea and control speaker could be utilized in two complementary ways: 1) statistical models trained on control (or healthy) speech, when used to test acoustic features coming from apnoea speakers should provide lower likelihood scores (i.e., control models will be "less likely" to generate apnoea speech due to OSA-related anomalies) than when testing control speakers (regarded that a consistent cross-validation scheme is used); and 2) apnoea/control classification can be considered as a speaker recognition problem using only two different statistical models, one trained for the apnoea group and the other for the control population. In this research we will explore the first way, as the second one has been considered in our previous work (Fernández et al. 2009).

5.1 Acoustic features

The **front-end** in any speech recognition system is the process involved in extracting a set of acoustic features from the speech signal, so that it provides an efficient representation of speech without losing discriminative information. These acoustic features should also correspond to the assumptions made by the actual modelling techniques (generally statistical independence between features). Selecting a proper parameterization is therefore a relevant task, and one that depends significantly on the specific problem we are dealing with. According to Fox and Monoson's (1989) perceptual experiments, some abnormalities can be directly identified by listening to the recordings. Therefore, conventional MFCC (Mel-Frequency Cepstral Coefficients) parameterization was applied in this research as it provides both, relative independent coefficients, and high discrimination between sounds based on its similarity with human perception processing (Huang et al. 2001). We acknowledge that an optimized representation, similar to that of Godino et al. (2006) for laryngeal pathology detection, could produce better results in terms of classification efficiency, but for the present work, we are not focusing on maximizing the accuracy rate, but in exploring differences within the acoustic space according to the same **principia** described in the preceding perceptual experiments.

5.2 Speech segmentation

To train different statistical models for different acoustic or linguistic units, the acoustic feature vectors resulting from the front-end pre-processing must be segmented or grouped into different training sets. Since we are interested in studying

specific phonetic classes, all of the utterances in our OSA database had to be segmented into phonetic units. This phonetic segmentation allowed us to group acoustic feature vectors with specific phonetic classes, and then to train a specific statistical model for each phonetic class.

All sentences in our apnoea database (both for control and apnoea speakers) were automatically segmented into phonemes through forced recognition. That is, each sentence was forced to be recognized using the sequence of phonemes corresponding to its known transcription (optional silences between words were allowed). This forced alignment provided the start and ending time boundaries for each sound in the sentence. Automatic forced alignment avoids the need for time-consuming and costly manual annotation, but, as will be discussed in Section 6, it must guarantee an appropriate level of segmentation precision. In our case automatic phonetic segmentation was carried out with the open-source HTK tool (Young 2002). We use 24 left-to-right, 3-state, context-independent **Hidden Markov Models** (HMMs) to represent the basic set of 24 Spanish phonemes. These context-independent HMM phoneme models were trained from an available manually segmented, phonetically-balanced speech subcorpus of Albayzin, a reference large speech database for Spanish (Moreno et al. 1993).

5.3 Statistical modelling

After phonetic segmentation, due to the fact that Mel-Frequency Cepstral Coefficients may follow any statistical distribution on different phonetic classes, the **Gaussian Mixture Model** (GMM) approach, broadly applied in speaker recognition systems (Reynolds et al. 2000), was chosen to approximate the actual statistical distribution of the selected acoustic space. In our experimental setup we started by training GMM models for different phonetic classes using a large speech database: the Albayzin database (Moreno et al. 1993). By doing so we provide a set of stable initial models from which, using adaptation techniques, more specific GMMs were derived (tuned to particular characteristics of the speakers' population, recording conditions, etc.). A MAP (*Maximum A Posteriori*) adaptation algorithm, also commonly used in speaker verification (Reynolds et al. 2000), was applied to derive those specific GMMs representing our OSA database peculiarities: limited in the amount of speech and more specific in their phonetic and population coverage. Additionally, MAP adaptation is known to increase the robustness of the models, especially when sparse speech material is available. Besides, as it is also a common practice in speaker verification systems, only the means of the gaussian components in the GMMs were adapted. For our experiments, MAP adaptation to GMM models was estimated with the BECARS open source tool (Blouet et al. 2004).

6 Modelling phonetic classes for OSA analysis

The basic unit to convey linguistic meaning is the phoneme. Each phoneme can be considered to be a code that consists of a unique set of articulatory gestures, which includes the type and location of sound excitation, as well as the position of the vocal tract articulators. Additionally, other factors, such as the resonances

produced within the vocal tract and the response of the vocal folds decisively affect the way in which those phonemes are pronounced. However, in this work we are not interested in the meaning, but in exploring the acoustic information embedded within speech signals. Consequently, we are not subjected to the traditional approach followed in speech recognition and may choose any other unit.

While specific instances of the individual phonemes are quite limited within short segments of direct speech, phonetic classes are easier to recognise than phonemes and occur much more frequently. Therefore models are easier to train with sparse data, as long as their internal complexity can be accommodated. Consequently a limited number of models can be trained when only sparse data are available. On the other hand, according to the previous literature dealing with the effects of OSA in speech signals, only a few phonetic classes seem to be relevant for our experiments. Bearing this in mind, four different groups of broad phonetic classes were defined:

Vowel sounds, VOW: vowel sounds represent one of the most relevant acoustic groups in speech processing applications, and have been intensively analyzed in the detection of pathological voices. Sustained vowel sounds typically are considered to be the best source of information. However, recent studies have pointed out that, at least for certain pathologies, vowel segments extracted from continuous speech might be as informative as those from sustained speech.

Nasal sounds, NAS: nasal sounds are especially relevant when considering resonance effects in speech signals involving both the oral and nasal cavities. The coupling and de-coupling of the nasal cavity, by means of the opening/closing of the velopharyngeal port, causes the most familiar resonance effect in speech. Nasal phonemes appear in conjunction with at least one vowel, and cause a singular unique transition from the vowel to the nasal (and vice versa) known as **nasalization.** This seems to be a particularly relevant situation (Davidson 2003), which we will be looking thoroughly at this paper.

Plosive sounds, PLO: in contrast to the two previous classes, plosive sounds represent non-stationary, fast transitions in the speech signal. Therefore, instead of cepstral coefficients, more specific acoustic measures (mainly voice-onset-time) are generally used for their study. Consequently, in our statistical models, built on cepstral coefficients information, plosive sounds could present lower variability rates. This is in contrast to vowel and nasal sounds, which are expected to exhibit variability when healthy and apnoea speakers are compared. However, due to co-articulation, and the flawed boundaries provided by our automatic segmentation process, the GMM model for plosive sounds could include acoustic information from transitions from adjacent phonetic classes. This could cause some differences in this class, when used as phonetic classifiers and thus become relevant to our research on apnoea speech.

Fricative sounds, FRI: an extra phonetic class is introduced in order to group all sounds which were not assigned to the previous classes. Considering our designed apnoea corpus, most of these sounds are fricative, although others, such as liquid sounds, will also be included in this fourth class. By grouping all of these sounds,

we complete our classification of sounds, introducing a quite artificial group which includes sounds with rather different characteristics, though, as we just said, fricatives form the most significant subset.

Using these four phonetic classes, our purpose will now be to explore any differences that could be found between OSA and healthy speakers using speech recognition technology. But before that, we have to give some details on how GMM models, as described in Section 5, were trained and how differences between phonetic classes were measured.

6.1 Training data and GMM characteristics

Considering the previous description of the four phonetic classes, it is important to note that as we are modelling them using GMMs, the linguistic differences between phonetic classes will not generate non-confusable or non-overlapping models. Besides the overlapping of the acoustic spaces in particular realizations of each phonetic class, the discriminative power of GMM-s depends on different factors, such as the size of the model (i.e., number of gaussians), amount of training data and the acoustic front-end parameterization. In our case, the automatic segmentation of phonetic units can also be a source or errors that, as we discussed before, could lead to some overlap between the acoustic spaces modelled by different phonetic-class GMMs. Being aware of all of these differences from ideal acoustic models, the use of broad phonetic classes allows us to ensure that, as long as our segmentation of the utterances is precise enough, the number of spurious frames will be negligible compared to the amount of reliable data, so little distortion is expected in the estimation of acoustic parameters. As we will see in Section 7, the trained GMM models deliver a classification rate that is accurate enough not only to discriminate between phonetic classes, but also to measure differences in the acoustic realizations between OSA and control speakers.

In summary, the full acoustic space in our speech database was divided, through automatic phonetic segmentation, into the four phonetic classes previously described (see top of Figure 1). Consequently, the amount of data available to train each of the four phonetic classes was different as well as the internal complexity of their statistical distributions. However, as the speech corpus was designed to have a homogeneous coverage of main phonetic contexts relevant to OSA pathology, we decided to model each phonetic class using GMM models with equal number of gaussian components. So, based on the amount of available training data, 64 gaussians were considered enough to properly represent the different acoustic complexities of the different phonetic classes.

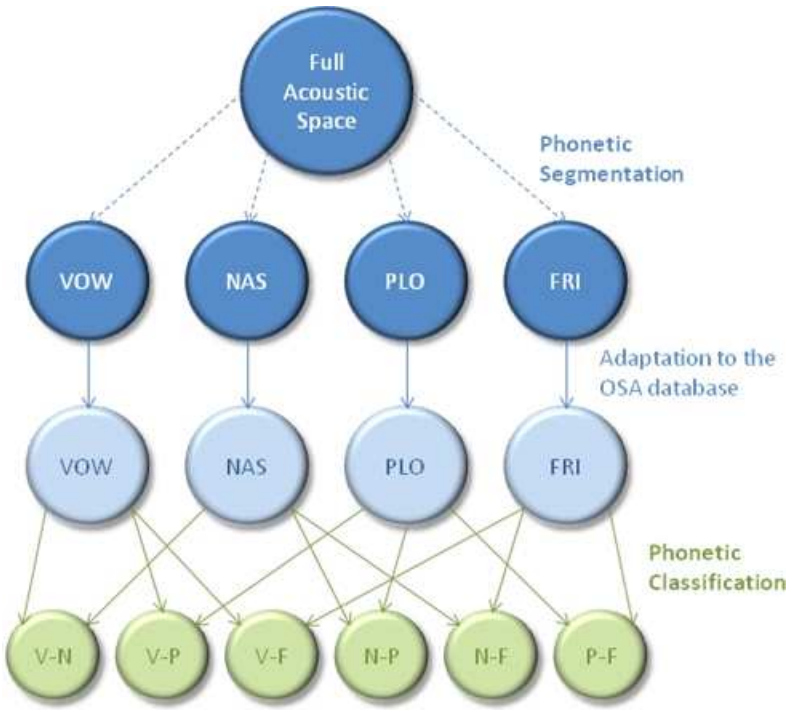


Figure 1. Brief description of the segmentation and adaptation processes (Section 6), as well as of the classification tests performed (Section 7)

6.2 Training reference GMMs for each phonetic class

As it was stated in Section 5, our aim is to explore acoustic differences between apnoea and control speakers using a set of GMMs trained from control or healthy voices. These voices served as reference GMMs to measure possible deviations of the **pathological voices** of apnoea speakers. As a result, these reference GMM models had to be trained from a control population, but, due to the limited amount of speech obtained from the control speakers in our OSA database, the first set of GMM models were trained using Albayzin database (Moreno et al. 1993). After that, as Figure 1 illustrates, these initial models were adapted to the control speaker population of our OSA database to generate the final four reference GMMs.

The whole training process can be described as follows: utterances from the Albayzin corpus, already manually segmented into phonemes, were labelled according to the four phonetic classes we defined. Once grouped, the feature vectors were used to estimate a GMM model for each phonetic class separately, resulting in four different models, namely: VOW, NAS, PLO and FRI. In the second step, as Figure 1 illustrates, a MAP algorithm (Reynolds et al. 2000; incorporating these initial GMM models) was used to adapt these models to the OSA database; specifically to its acoustic conditions (microphone and recording room) and population characteristics. To complete this adaptation process, speech utterances from the control speakers were automatically segmented and labelled using the

process described in Section 5. Note that, as it was said before, only control speakers were used in the adaptation process to generate the final set of reference (healthy) GMM models for each phonetic class.

6.3 Measuring differences between phonetic classes

Once reference GMM models were trained, we tested their ability to classify different speech segments as belonging to different phonetic classes. Our hypothesis was that their discriminative power will be lower for apnoea speakers than for control speakers, as the pathological characteristics of apnoea voice make them more confusable. These differences in phonetic class discrimination then could be exploited to detect apnoea cases.

Thus, given a set of acoustic features corresponding to a particular phonetic class produced by a particular speaker, the discriminative measure used will be the difference between the logarithm of two likelihood scores (i.e. log-likelihood ratio, LLR, Reynolds et al. 2000): one score was the log-likelihood of generating the set of acoustic features using the true GMM model (i.e., the correct phonetic class), and the other score was the log-likelihood obtained using a different phonetic class. As can be seen in the bottom of the diagram in Figure 1, we explored the differences between every pair of GMM phonetic classes (V-N, V-P, V-F, N-P…).

We should bare in mind that the difference between likelihood scores are closely related to the Kullback-Leibler divergence (KLD) –also known as *discrimination information*– [17]. KLD is the most common approach to measure differences between the statistical distributions of two classes and to decide which of the two models most likely generated a certain sample. This theoretical measure can be estimated either by calculating the average likelihood ratio between two models over a set of feature vectors, or by considering an analytic approximation to it. This analytic approach will only be used in subsection 7.4, while likelihood averaging will be used in subsections 7.1 to 7.3.

7 Experimental results

Several experiments were developed to explore differences between the four phonetic classes, and all of them were based on the differences between log-likelihood ratios (LLRs) for control and apnoea speakers. To provide a fair test, both the adaptation of the reference GMM models and LLRs were estimated using the leave-one-out cross-validation test protocol. According to it, for all tests involving a particular speaker in the control group, the four reference GMMs were trained through MAP adaption using our OSA database, but excluding (leaving-out) this particular speaker's records. Z-score normalization was used to fairly compare results for the different phonetic classes and to consider their posterior fusion at the score level.

To quantify the acoustic mismatch between apnoea and control speech, two different approaches were considered, both of which were evaluated over a given sequence of acoustic features belonging to a particular phonetic class:

• First, our reference GMM models were used as classifiers of phonetic classes. We explored whether different performance rates in classification could be found for the control and apnoea populations. Note that in this case, we were not really classifying control/apnoea speakers but only exploring whether significant differences in phoneme classification exists across both groups. This result will provide some insights into the effects of apnoea on the speech of OSA patients.

• In a second set of experiments, control/apnoea classification was evaluated using average LLR from the reference GMM model corresponding to the true phonetic class, and the GMM of a different or competing phonetic class. Due to voice anomalies in apnoea patients, this average LLR was found to be different for control and apnoea speakers (i.e. higher for control speakers and lower –greater confusability– for apnoea speakers).

Finally we will conclude this Section by discussing how GMM models trained for vowel sounds in nasal and non-nasal contexts show an interesting distinctive pattern for apnoea speakers that should be explored in future research.

7.1 Differences in classifying phonetic classes

In this initial experiment, the discriminative power of the reference GMM models were evaluated using them for classification and comparing them across both the apnoea and control populations.

For each speaker in our database all the speech segments corresponding to the different phonetic classes were used to obtain the average LLR scores. Those were calculated as the mean difference of the log-likelihood values estimated for each speech sample by considering two reference GMM models (each of them corresponding to a phonetic class model). Thus, for each pair of phonetic classes, two different errors were possible: a) **missed recognition,** when a speech sample belonging to the first class was more likely to be generated by the second one, and b) a **false alarm,** when a speech sample belonging to the second phonetic class was more likely to have been generated by the first phonetic class model being evaluated. Depending on the decision threshold used across LLR scores, these two types of errors should be **opposite** (i.e. lower false alarm rates lead to higher missed recognitions, and vice versa) and can providing different operational points. Detection error trade-off (DET) curves have been widely used to represent the evolution for both types of errors (Reynolds et al. 2000), but also the discriminative power of a classifier can be described using a single Equal Error Rate value (EER). The EER corresponds to the operational point of equal missed recognition and false alarm errors. In Table 2, EER values representing the pair-wise phoneme class classification errors using the reference GMM models are presented. Different EER values are presented for both control and apnoea (bold values) populations.

From these results, we can see that classification rates are significantly different from one class to the other, though the results are reasonably good for all of them (the worst case being an EER of 11.7% classifying plosives vs. fricatives in the apnoea population). For vowels, results were particularly good when compared to those for nasals and plosives, as almost no errors appeared when testing over the

whole data set for both groups of speakers. Other pairs do exhibit small, but meaningful, error values with quite different results. However, they all reflect a common trend: performance for the apnoea group is worse than the control group. This result suggest a systematic deviation from the reference acoustic phonetic classes in apnoea speakers which can be related to the physiological factors associated with OSA. For instance, the increase in EER value when comparing nasal sounds (NAS) to fricatives (FRI) and plosives (PLO) can be explained by the fact that patients suffering from the OSA syndrome exhibit abnormal velopharyngeal function, so this could alter the production of nasal sounds, introducing a slight oral plosive and fricative articulation due to partial palatal paralysis.

Table 2. EER values resulting from phonetic classification of all pairs for the four phonetic classes. Bold values correspond to the apnoea population, while the normal ones were estimated for the control group.

	VOW	NAS	PLO	FRI
VOW		0.0% **0.0%**	0.0% **3.3%**	1.7% **3.3%**
NAS	–		0.0% **4.2%**	1.7% **6.7%**
PLO	–	–		3.3% **11.7%**
FRI	–	–	–	

7.2 Phonetic classes for OSA detection

Based on the different classification results for control and apnoea populations previously described, we will now analyze whether the underlying differences in LLR scores could be used to classify a speaker as belonging to the control or apnoea population. LLR was evaluated in the same way as described in subsection 7.1: using two competing GMM models, but in this case, it was only averaged for speech segments corresponding to a single phonetic class. That is, in this experiment the phonetic class of the speech segment was known (as provided by the automatic phonetic segmentation process), but whether the speaker belonged to control/apnoea group was unknown.

Therefore, for a given speaker to be tested, 4 speech segments, one for each phonetic class, were used, and, for each segment, 3 different average LLRs were obtained. For example, for the speech segments corresponding to the vowel phonetic class, three different LLR scores were obtained using the V-N, V-P and V-F pairs of reference GMM models. Consequently, using each one of these three LLRs, three different control/apnoea classification results were considered. So far, when speech segments for all phonetic classes were used, and LLRs for all possible combinations of reference GMMs were used, a total of 12 control/apnoea classifiers were evaluated.

As in the previous experiment, evaluation for this set of control/apnoea classification systems was based on the **miss recognition** and **false alarm** errors, but in this case missed recognition meant that an apnoea speaker was incorrectly

classified as a control speaker, and a false alarm signalled a control speaker being classified as having apnoea. Control/apnoea classification results, in terms of EER values for each one of the 12 classifiers, are presented in Table 3.

Table 3. EER values for control/apnoea classification using speech segments of the four phonetic classes and LLR scores for all pairs of reference GMM models

	VOW	NAS	PLO	FRI
VOW-NAS	46.7%	38.3%	–	–
VOW-PLO	42.5%	–	47.5%	–
VOW-FRI	47.5%	–	–	40.8%
NAS-PLO	–	37.5%	50.0%	–
NAS-FRI	–	39.2%	–	44.2%
PLO-FRI	–	–	46.7%	33.3%

From these results we can see that apnoea could be detected with an accuracy as high as 33% EER, which is rather surprising as this best classification result was obtained when considering fricative samples evaluated using LLR scores from fricative vs. plosive reference GMMs. In contrast, a very poor discrimination rate was attained when plosives were compared to fricatives. A possible explanation for this apparently odd result could be that in this experiment what we consider is not just the deviation from a perfect fit to the reference phonetic class models, but also the deviation towards a certain phonetic class. So in this case, fricative sounds in the apnoea group show a deviation towards plosive reference sounds. The same idea explains the results obtained when we compared nasals and vowels or nasals and plosives. Looking at other results in the Table, there are cases where both comparisons provided rather similar results for samples from both classes. This finding indicates that the distortion in one direction is about the same in the opposite one, just as it happens for vowels and plosives, vowels and fricatives or nasals and fricatives.

The results from nasal speech segments (NAS column in Table 3) require a more extensive explanation. According to the reviewed literature, abnormal resonances in speech are characteristic of OSA patients, particularly when considering the nasalization of connected vowels. Therefore, it was expected that nasal sounds would be useful cues in the design of an automatic system for OSA detection. In fact, Table 3 shows lower global EER values for the NAS column when compared to other phonetic classes. Consequently, the effects of vowel nasalization required from a specific analysis, which we describe in Section 7.4 by considering two different phonetic class subsets for vowels: those in nasal or non-nasal phonetic contexts.

7.3 Improving detection by the combination of pairs

From Table 3, it seems clear that classification results are poor for each of the individual classifiers. In this section, we will try to improve those results by

combining all 12 classifiers into a single one. This is a complex task which generally requires a large amount of data to guarantee that the optimal combination is found. Since the current dataset is small, we could not implement an optimal approach, but used a suboptimal one, which was designed to iteratively improve binary classification.

The combination process used was based on the algorithm described by Al-Ani et al. (2003), though conditional mutual information calculations were substituted by EER estimations, which are in fact the posterior error probabilities discussed in that article. The idea was to improve classification rates by linearly weighting normalized scores and adding them up; but only if the overall results were noted to improve. In order to avoid any redundancies and spurious effects which could detrimentally affect the results, all combinations (successive pairs, triplets, quartets, etc.) were tested in order to identify the optimal one. However, as suggested by Al-Ani et al. (2003), good results (though suboptimal) can be obtained by iteratively combining the weighted classifier with the best and most uncorrelated spare classifier, reducing the computational complexity.

The results from all these combinations are presented in Figure 2 using DET curves. The final DET curve, corresponding to the combined system, returns a 28.33% EER. This final DET curve is presented along with a different set of DET curves in Figures 2a and 2b. In Figure 2b (right plot), the different successive DET curves illustrate how successive classification improvements are obtained during the iterative algorithm.

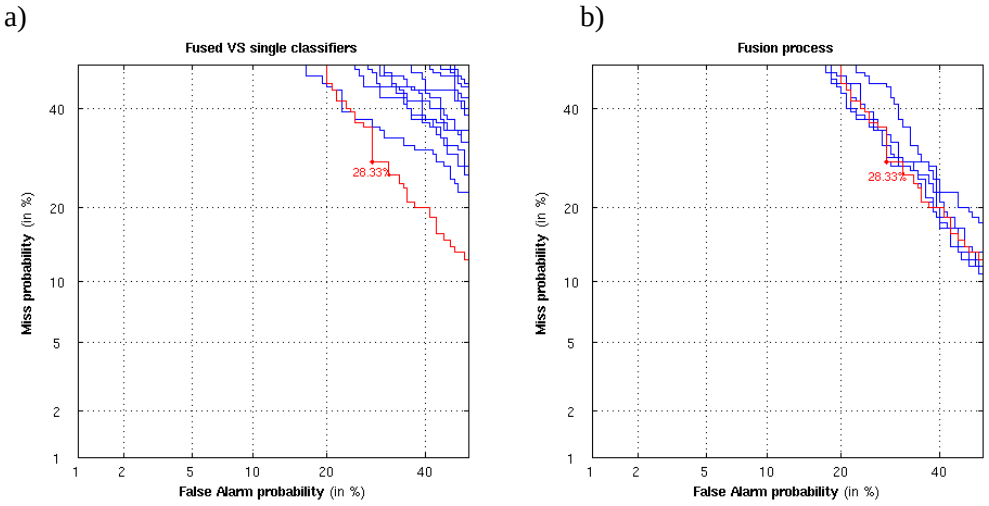


Figure 2. DET curve resulting from the combination of the 12 phonetic classifiers: the left one (a) compares the resulting DET curve with the ones estimated for each single classifier; the right one (b) compares the results from the iterative improvement algorithm.

As can be seen in this Figure, most of the improvement is achieved during the first iterations, while only marginal improvements occur later in the process. When comparing the final combination result to each of the twelve prior classifiers (the 12 upper right DET curves shown in Figure 2a), it becomes clear that there is a strong correlation among classifiers, although there is a considerable gap between the best prior classifier and the one resulting from the fusion process, improving the overall classification by 5%.

7.4 Exploring vowel nasalization

With the previous results estimated for control and OSA speakers in mind, nasalization effects (affecting both nasal and connected vowel sounds) seem to be a relevant phenomenon in apnoea detection. In order to improve our understanding of the side effects of the abnormal coupling and decoupling of the nasal and oral cavities, as well as to continue to rework Fox & Monoson's (1989) experiments by means of automatic speech processing, an additional exploratory experiment was carried out. The abnormal resonances described in Fox and Monoson's work could be perceived as a form of either hyponasality or hypernasality (no nasalization is produced when the sound should be nasal, or nasalization is produced during the pronunciation of non-nasal –voiced oral– sounds). In other words, OSA speakers will nasalize when they are not expected to, and/or vice versa. As a consequence, we will expect statistical models (GMMs) trained with such data to exhibit smaller differences when comparing models for vowel sounds in nasal and non-nasal contexts. This idea could be tested by measuring the distances between both models in each group of speakers.

Acoustic feature vectors for vowel sounds were grouped into two different subsets, based on whether their phonetic context was nasal or non-nasal, i.e. depending on whether they should be nasalized or not. The amount of available data for the original VOW phonetic class was enough to build the class model for the previous experiments, and is even big enough for our tests once we redistribute samples among these two nasal sub-classes. However, since we have reduced the size of the data set in this experiment, the KLD analytic approximation (Do 2003) was chosen. Therefore, four different models were trained by adapting the original VOW GMM: two GMMs adapted to vowels in non-nasal context (one for control and the other for apnoea speakers), and two GMMs for vowels in nasal contexts (also for control and apnoea voices).

As a test of the stability or consistency of our KLD approximation, these four GMM models were trained and the corresponding KLD distances were evaluated 40 times, each time using a different subset of 39 control and apnoea speakers extracted from our database. Figure 3 represents the resulting 40 KLD distance values obtained for GMM models for vowels in nasal and non-nasal contexts (speaker index in the Figure corresponds to the excluded speaker in the 39 speakers' subgroup). As it can be seen in Figure 3, significant differences in the nasal/non-nasal GMM distances were found for the control and apnoea speakers. This result suggests that acoustic differences between oral and nasal vowels are smaller in

apnoea speakers and confirms the trend to an overall higher nasality level, as revealed in previous research.

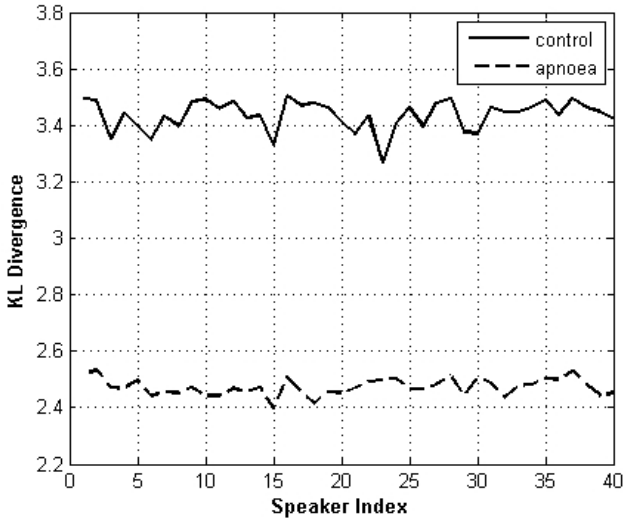


Figure 3. KLD approximation values between Gaussian Mixture Models for vowels in nasal and non-nasal contexts

8 Conclusions and future research

In this paper, some of the characteristic speech patterns that can be observed in speakers suffering from severe obstructive sleep apnoea (OSA) syndrome have been analyzed by comparing phonetic classes using a specifically designed speech corpus. This study offers an innovative perspective on how phonetic information can be used in pathological voices analysis using conventional automatic speech processing techniques.

Regarding Fox & Monoson's research as a reference, a "perceptual" representation of the speech signal using Mel-frequency cepstral coefficients was used. From this acoustic representation, experimental results were obtained using Gaussian Mixture Models (GMMs), which were initially trained on a large Spanish speech database and adapted to a control population. These GMMs were generated for the four broad phonetic classes and then used as reference patterns to explore possible acoustic mismatches in the voices of speakers with apnoea.

Differences in phonetic classification for control and apnoea populations were observed for the four phonetic classes. These results suggest that certain phonetic groups are more likely to be misclassified when the speaker suffers from apnoea. Using all different pair-wise reference GMM models, control/apnoea classification was also evaluated using log-likelihood ratio scores averaged over segments of speech corresponding to different phonetic classes. The minimum 33% EER

obtained when using single classifiers, was improved to 28.3% when combining all of them through an iterative linear weighting algorithm.

Finally, various effects addressed in the previous literature were identified in our experiments, supporting the interpretation of the automatic speech recognition results. Reworking Fox and Monoson's (1989) experiments has allowed us to come to the same conclusions they did. Though further analysis is needed, apnoea speakers certainly exhibit smaller intra-class differences during vowel nasalization. This side-effect is probably related to an abnormal coupling of the nasal cavity.

Our results are intended to shed some light on the peculiarities that phonetic classes exhibit when comparing healthy speakers to those suffering from OSA. Results obtained in control/apnoea classification were also promising, though still much work needs to be done. Besides, there is still a need for a larger speech database to continue study in this area. We shall focus on this need, while encouraging research to improve our understanding of the effects of OSA on speech signals.

9 Acknowledgments

The activities described in this paper were funded by the Spanish Ministry of Science and Technology as part of the TEC2009-14719-C02-02 project. The authors would like to thank the volunteers at Hospital Clínico Universitario of Málaga, Spain, and to Guillermo Portillo who made the speech and image data collection possible. José Luis Blanco also acknowledges Universidad Politécnica de Madrid scholarship's support.

References

Al-Ani, A., Deriche, M. and Chebil, J. 2003. A new mutual information based measure for feature selection. *Intelligent Data Analysis,* 7(1): 43-57.

Blouet, R., Mokbel, C., Mokbel, H., Sanchez Soto, E., Chollet, G. and Greige, H. 2004. BECARS: a free software for speaker verification. In *Proceedings of The Speaker and Language Recognition Workshop.* ODYSSEY. 145-148.

Coccagna, G., Pollini, A. and Provini, F. 2006. Cardiovascular disorders and obstructive sleep apnoea syndrome. *Clinical and Experimental Hypertension,* 28: 217-224.

Davidson, T. M. 2003. The great leap forward: The anatomic evolution of obstructive sleep apnoea. *Sleep Medicine,* 4: 185-94.

Do, M. N. 2003. Fast approximation of Kullback-Leibler distance for dependence trees and Hidden Markov Models. *IEEE Signal Processing Letter,* 10: 115-118.

Fernández, R., Blanco, J. L., Hernández, L. A., López, E., Alcázar, J. and Torre, D. T. 2009. Assessment of severe apnoea through voice analysis, automatic speech, and speaker recognition techniques. In *EURASIP Journal on Advances in Signal Processing,* vol. 2009. 1-21.

Fernández, R., Hernández, L. A., López, E., Alcázar, J., Portillo, G., and Torre, D. 2008. Design of a multimodal database for research on automatic detection of severe apnoea cases. In *Proceedings of 6th Language Resources and Evaluation Conference*. Marrakech, Morocco. 1785-1790.

Fox, A. W. and Monoson, P. K. 1989. Speech dysfunction of obstructive sleep apnea. A discriminant analysis of its descriptors. *Chest Journal,* 96(3): 589-595.

Godino-Llorente, J. I., Gómez-Vilda, P., Sáenz, N., Blanco-Velasco, M., Cruz, F. and Ferrer, M. A. 2006. Support vector machines applied to the detection of voice disorders. In Hussain, A., Faundez-Zanuy, M. and Kubin, G. (eds.): *Lecture notes in computer science, Vol. 3817: Nonlinear analyses and algorithms for speech processing.* Heidelberg: Springer Verlag. 219-230.

Hébert, M. and Heck, L. P. 2003. Phonetic class-based speaker verification. In *Proceedings of the Eurospeech 2003 Conference*. Geneva, Switzerland. 1665-1668.

Huang, X., Acero, A. and Hon, H-W. 2001. *Spoken language processing: A guide to theory, algorithm and system development.* New Jersey: Prentice-Hall.

Kullback, S. and Leibler, R. A. 1951. On information sufficiency. *Annals of Mathematical Statistics,* 22(1): 79-86.

Lee, W., Nagubadi, S., Kryger, M. H. and Mokhlesi, B. 2008. Epidemiology of obstructive sleep apnea: a population-based perspective. *Expert Review of Respiratory Medicine,* 2(3): 349-364.

Lloberes, P., Levy, G., Descals, C. et al. 2000. Self-reported sleepiness while driving as a risk factor for traffic accidents in patients with obstructive sleep apnoea syndrome and in non-apnoeic snorers. *Respiratory Medicine,* 94: 971-976.

Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J. B. and Naude, C. 1993. ALBAYZIN speech database: Design of the phonetic corpus. In *Proceedings of Eurospecch 93*. Vol. 1. Berlin, Germany. 175-178.

Reynolds, D. A., Quatieri, T. F. and Dunn, R. B. 2000. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing,* 10: 19-41.

Robb, M., Yates, J. and Morgan, E. 1997. Vocal tract resonance characteristics of adults with obstructive sleep apnea. *Acta Otolaryngologica,* 117: 760-763.

Turner, G. S., Tjaden, K. and Weismer, G. 1995. The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis. *Journal of Speech and Hearing Research,* 38: 1001-1013.

Young, S. 2002. *The HTK Book (for HTK Version 3.2).* First published December 1995, Revised for HTK Version 3.2 December 2002.

LABORATORIES OF SPEECH RESEARCH & TECHNOLOGY AT THE DEPARTMENT OF TELECOMMUNICATIONS AND MEDIA INFORMATICS (TMIT), BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS (BME)

http://www.tmit.bme.hu

History of the Laboratories

Prof. Géza Gordos, founding Head of the Department, started to build a speech research and technology laboratory at the Budapest University of Technology and Economics in 1969. Since then, this laboratory has ever since been active in practically every aspect of speech processing.

BME TMIT has been represented at all biennial Eurospeech conferences since 1987. In recognition of our achievements, the ESCA board (European Speech Communication Association, which today is known as the International Speech Communication Association, ISCA) selected this institute to host Eurospeech '99 with Prof. Géza Gordos as General Chairman, Prof. Mária Gósy as Co-Chairman and Dr. Géza Németh as the Scientific Secretary (app. 730 papers were presented, with 1100 participants from 51 countries).

Members of BME TMIT maintain both strong academic and industrial contacts including national and international projects (i.e., COST, ACCORD, Copernicus and EU R&D Framework). Our various results have been utilized in the industry since 1982. The MULTIVOX multilingual text-to-speech (TTS) system – developed in our department – was one of the Hungarian winners of the Software for Europe competition. Also in 1996, a measuring instrument developed in our Telecommunications and Signal Processing Laboratory received the European Innovation Award and the Innovation Award of the Hungarian Ministry of Telecommunications and Transportation. Five current (and one previous) researchers from BME TMIT (Géza Gordos, Géza Németh, Gábor Olaszy, Péter Tatai, György Takács and Klára Vicsi) received the shared Award of the Hungarian Academy of Sciences for outstanding results in speech communication and technology in 1999. R&D results are taken into the classroom coursework, laboratory work, diploma theses and PhD training. The Speech Information Systems aspect of the department attracted 95 students in its first year of operation in 2001, which then increased to 311 in 2006.

The present

Speech research and technology-related activities are presently conducted within the Section of Telecommunications and Speech Systems of BME TMIT, which is headed by Prof. Géza Gordos. The team has approximately 40 researchers and students. In 2004, a spinoff company for research was created (http://www.aitia.ai).

Laboratories of the Section contributing to this activity and their major fields of interest are:

Speech Technology Laboratory (http://speechlab.tmit.bme.hu)

Contact: Dr. Géza Németh (nemeth@tmit.bme.hu)

Multilingual text-to-speech synthesis

• MULTIVOX text-to-speech system (multilingual grapheme-sound conversion, prosody modelling, formant synthesis, supporting 10 languages, 1986-1996).

• PROFIVOX waveform based TTS development environment (implemented for Hungarian, German, Polish and Spanish, with client-server architecture, TTS high-level control mark-up language MVML, description of detailed data and rule systems for Hungarian TTS and automatic prosody generation (timing rules, amplitude structures and multilevel F0 rules) are available) (1994-).

• High quality number-to-speech generation (implemented in Hungarian, German and English, starting in 1996 to the present).

Computer telephony integration (CTI), dialogue systems

• Hungarian person, company name and address reader (2000-04, operated by T-Mobile Hungary as a reverse directory application).

• World's 1st Symbian mobile phone-based SMSreader product (2003-present, together with M.I.T. Systems Ltd., marketed in Hungary by T-Mobile Hungary, English product name: SMSrapper).

• 1st Hungarian VoiceXML browser (2002-03, based on OpenVXI).

• Development of the first multi-line, network based Hungarian e-mail and SMS reading system with automatic diacritic regeneration (1998-present).

• Development of commercially used audiotext/voice response applications, including the first Hungarian Speaking Bill over the Telephone (1995) and the first Hungarian residential voice-mail system (1996).

• Development of the speech interface of an automatic announcement system for the Hungarian Telecommunications Company used on app. 600.000 lines, based on formant synthesis (1992).

• Human factors, use ability issues (COST219, EU project Mobile Rescue Phone, 1990-).

Applications of the synthesis engine for the disabled and the elderly

• Speaking systems included in screen readers for the blind (e.g., Hungarian Jaws for Windows) and the speech impaired (1984-).

Laboratory of Speech Acoustics (http://alpha.tmit.bme.hu/speech/)

Contact: Dr. Klára Vicsi (vicsi@tmit.bme.hu)

Database collection

• SpeechDat(E): telephone speech database, a realistic speech base, both for the realistic training and testing of present-day teleservices and for the training of real speaker independent speech recognizers, European project (1999-present).

• BABEL - a multilingual speech database collection clear read speech for general speech processing purposes of (1995-99) INCO - COPERNICUS project.

• CHILDREN SPEECH data-base collection (1999-present). Part of the SPECO Copernicus program.

• MTBA, BESZTEL-Speech database collection through telephone lines and mobil telephones (2000-2003).

Annotation and segmentation

• Automatic speech segmentation on phonetic, sub-phonetic level of continuous speech, automatic labelling (1997-present).

• Language independent automatic segmentation and labelling technique has been developed for training speech recognizers, specifically to collect base (1990-present).

Speech recognition

• Speaker independent, isolated word robust (noisy telephone) speech recogniser (1989-present).

• Neural network based speech recogniser, supported by phoneme, diphone, half-syllable based recognition on phonetic and phonological level (1995-present).

• HMM based middle sized, speaker independent, continuous speech recognition for fixed topics (2004-present).

Speech processing for speech-pronunciation teaching systems for speech handicapped and for language learning

• Audio-visual speech-pronunciation teaching system for individuals with speech impairments and for pronunciation training in language learning.

• A Multilingual pronunciation teaching and training method within the EU Copernicus program, entitled SPECO.

Statistical examination of the Hungarian language

• Search of optimal units of CSR systems.

• Construction of optimally sized teaching and testing material, based on the nature of the language concerned.

• Language modelling.

Telecommunications and Signal Processing Laboratory (http://ds.aitia.ai)

Contact: Péter Tatai (tatai@tmit.bme.hu)

Automatic Speech Recognition

• Basic research: exploring and solving problems related to the Hungarian language, which is highly agglutinative, and therefore most words have hundreds or even more different forms, making word based recognizers generally impractical.

• Speaker-independent open vocabulary speech recognition-based telephone information services (open vocabulary: vocabulary extension by text input).

• Realization of voice controlled call centers and web based voice portals.

• Automatic speech recognition based call center: “Voxenter”.

• Finite state grammar- constrained, connected-word recognition, using HMM technology.

• Development of a system for automatic prediction of all possible pronunciations of Hungarian words.

• Development of a morpheme-based grammar-model for the recognition of all inflected forms of Hungarian words.

Speech quality measurement

Speech quality measurement system including

• subjective testing tools - absolute and comparison tests,

• objective testing - calibrated to the subjective results,

• S detection and other special signals in the channel,

• on line observation of the transmission for calibration.

Front-end development

• Enhanced line spectrum estimation, cepstral trajectory approximation with FFT for automatically segmented subword units, standard front-ends are also available.

• Automatic subword (demi-syllable) segmentation.

Language modelling

• FSA grammar descriptions for specific applications.

• Written text to phoneme sentence conversion.

• Triphone set creation, taking coarticulation effects into account (manual - automatic).

• Pronunciation training database - transcription of 1.5 million Hungarian word forms.

Géza Németh
Budapest University of Technology and Economics
Budapest, Hungary
e-mail: nemeth@tmit.bme.hu

INTEGRATING PROSODY INTO AUTOMATIC SPEECH RECOGNITION

György Szaszák
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics, Hungary
e-mail: szaszak@tmit.bme.hu

Abstract

This paper is a short overview of the author's dissertation titled "The role and usage of supra-segmental features in automatic speech recognition". Generally prosody is neglected in speech recognition, although a part of the information transmitted by speech is encoded in form of prosodic features. This paper presents the possibilities of acoustic-prosodic preprocessing and some methods which allow for the integration of prosodic information into the automatic speech recognition process. Prosody based speech segmentation and automatic sentence type recognition are presented and evaluated.

1 Introduction

In the age of electronic information and artificial intelligence, machines are increasingly expected to implement or extend human's capabilities. The same is true for automatic speech recognition: humans would like to communicate with computers using a natural language. Since the beginning of speech recognition research, several revolutionary innovations have contributed to the improvement of automatic speech recognition, however, there is still much to do.

Standard speech recognition focuses on the transcription of the speech. This is a pure phone-to-grapheme conversion based on spectral features extracted in the segmental domain. Any additional information source is discarded and not processed by standard automatic speech recognisers (ASR), like the prosody of speech which is important in decoding sentence type, modality or emotions. ASR systems interpret the speech as a sequence of phonemes, and speech processing is carried out strictly in the segmental domain (Jelinek 1998): the acoustic-phonetic level in speech recognition is modelled by the so-called acoustic phoneme models, and the acoustic speech signal is used only to obtain phoneme sequence hypotheses. Phoneme sequences then are grouped into words, which are managed by the language model in order to statistically specify how words can form the word chains that are the output of most recognizers. The author argues that supra-segmentals acoustically support the word chain level mentioned above and can help to develop more robust speech recognition systems. In addition, a speech utterance with improper prosodic structure (improper accents, stress or intonation; too monotone speech) degrades the performance of human speech understanding.

As far as the author knows, applications in which prosodic modules are implemented are rarely used in automatic speech processing; an exception is the prosodically rich, but not tonal Japanese language, where the fundamental frequency contours of words or parts of words have been modelled and recognized (Hirose 2001). Unfortunately, these results are presented only for a test set composed of only two speakers, however, the results were impressive with an improved mora recognition rate noted. For American English, a decision tree-based system was published (Veilleux & Ostendorf 1993), which uses fundamental frequencies and speech breaks to reorder the N-best hypotheses obtained from the speech recognizer. In this case, a syntactic and semantic analyser was used to create a reference prosodic structure for each hypothesis. Hypotheses were then rescored based on a coincidence evaluation between the generated and observed prosodic structure. The primary prosodic feature utilized was duration, as it plays a major role in American English speech prosody, including stress. A similar approach also was published for the German language (Kompe 1997; Kompe et al. 1995). For Standard Colloquial Bengali, a language spoken in India, a word boundary detector has been developed using prosodic features (Mandal et al. 2007). Detection of prosodic phrase boundaries and sentences, or speaker segmentation are relatively often-published applications of speech prosody (Cristophe et al. 2004; Kompe 1997; Shriberg et al. 2000; Waibel 1988).

In the present article, which provides an overview of the authors' dissertation (Szaszák 2009) the role of prosody is addressed with a special focus on fixed-stress languages. The goal was to perform automatic speech segmentation for units shorter than prosodic phrases. It will be shown that such segmentation can improve automatic speech recognition performance. Finally, the automatic detection of the sentence type based on prosody also will be presented.

2 Subjects, material and method

The speech resources used were the Hungarian BABEL database (Vicsi & Vig 1998: the part used contained about 1600 sentences from 32 speakers), Hungarian Children Database (approx. 18000 utterances), the Finnish Speech Database (Vainio et al. 1999: 250 sentences from 4 speakers) and a portion of the German Kiel Corpus (tales "Nordwind und Sonne" and several individual phrases; *KIEL Corpus of read Speech* 1994). These databases were manually pre-segmented into prosodic phrases and clauses/sentences.

Two main tasks were conducted: **speech segmentation** based on prosody (for Hungarian and Finnish), and **sentence type recognition** (for Hungarian and German). The prosodic attributes used were fundamental frequency and energy, as they can be easily extracted from the speech signal automatically. Duration measures were not used, because the author found during his preliminary experiments (Vicsi & Szaszák 2005) that they were not consistently used in the Hungarian speech database, and because their extraction would be too time-consuming (i.e., phone-level segmentation would be required). The fundamental

frequency and energy contours were extracted using the Snack toolkit and then processed and smoothed to correspond to the supra-segmental domain.

The segmentation of speech was based on the presence of stress in the F0 and/or energy contours, which were automatically detected by either peak-detection algorithms or by hidden Markov-model (HMM) based alignment. Both approaches assume that the stress in the language is fixed. If a HMM-based alignment is used, the prosodic contours of phonological phrases (Hunyadi 2002; Varga 2000) are modelled based on their F0 and energy contours. For a fixed-stress language (Hungarian and Finnish are fixed-stress), the stress within the phonological phrase has also a relatively fixed position. Six different prosodic contours were modelled, including silence. Experiments were carried out for Hungarian and Finnish in order to investigate word-boundary detection based on phonological phrase alignment (a phonological phrase was presumed to correspond to distinct words or word-chains and hence, phonological phrase boundaries were regarded also as word boundaries). The HMM models utilized were 11-state left-to-right ones with 1 or 2 Gaussians per state.

The performance of the algorithms was evaluated using two measures, **precision** *(p)*:

$$p = \frac{tp}{tp + fp} \cdot 100\% \tag{1}$$

where tp (true positive) stands for the number of correctly detected word-boundaries, fp (false positive) is the number of false detections. The measure **recall** (r) refers to the ratio of detected word boundaries from all word-boundaries:

$$r = \frac{tp}{tp + fn} \cdot 100\% \tag{2}$$

where fn (false negative) stands for missed (not detected) word-boundaries.

The detected word-boundaries were then used to perform N-best rescoring (Veilleux & Ostendorf 1993) in ASR in order to investigate whether word-boundary information improves speech recognition performance (i.e., reduces the error rate at the output of the ASR). This was investigated only for the Hungarian language.

For automatic sentence type recognition, the same HMM framework was used as for phonological phrase alignment and derived word-boundary detection: HMM models were trained to match the specific prosodic contour of clauses depending on the type of the sentence. A silence and a non-terminal contour model was also used in addition to the sentence type-specific terminal contour models. Hungarian and German languages were utilized during these experiments For Hungarian, seven different contours were modelled and for German five contours were developed (because less data was available). To evaluate the implemented system, the ratio of correctly recognized clause units/sentence types was measured:

$$Corr = \frac{H}{N} \cdot 100\% \tag{3}$$

where *H* is the number of correctly recognized clauses/sentence types and *N* is the total number of clauses/sentence types in reference transcription.

3 Results

3.1 Automatic classification of phonological phrases and word-boundary detection

The modelled phonological phrase contours were (Szaszák 2009): falling (FA), descending (DE), floating (FL), rise-fall (RF), ascending/rise (RI), and silence (SIL). All phonological phrases except floating and silence contours were expected to start with a stressed syllable and continue with a specific intonation contour (i.e., the phonological phrase is interpreted as a prosodic unit with distinct stress). The boundaries of phonological phrases always coincided with word-boundaries (but the reverse was not always true). In this case, the Hungarian and Finnish languages were involved in several training-testing conditions, as shown in Table 1. Word-boundary detection results were recorded and the results are shown in Figure 1.

Table 1. Training and testing configurations of phonological phrase alignment based word-boundary detection

Training	Testing	Code
Hungarian	Hungarian	HH
Hungarian	Finnish	HF
Finnish	Hungarian	FH
Finnish	Finnish	FF
both Mixed	Hungarian	MH
both Mixed	Finnish	MF

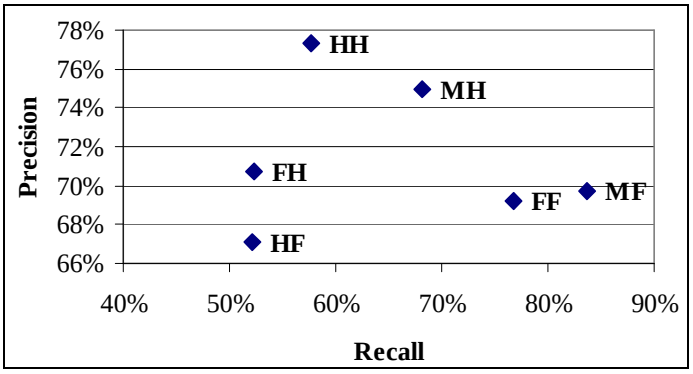


Figure 1. Precision and recall of word-boundary detection for Hungarian, Finnish and mixed systems

3.2 Improvement of ASR

To analyze the effect of word-boundary information on ASR performance, the Hungarian HMM-based phonological phrase alignment system was used for word-

boundary detection, and the detected word boundaries were used to rescore N-best lattices in a Hungarian medical ASR application (see Szaszák 2009 for details). The N–best rescoring algorithm augments the scores of word-chain candidates if the word boundaries match well the detected ones. In case of considerably mismatch between the detected (phonological phrase alignment) and hypothesized (speech recognition procedure) word-boundaries, the score is reduced. The speech recognition procedure was split into two parts: in the first part, the N–best lattice was obtained, which is then rescored based on the prosodic segmentation given by word-boundaries. Finally, in the second part of the recognition process, the lattice was parsed to obtain the recognition result.

The experimental testing set was composed of 20 medical reports (approx. 300 sentences in all). Speech recognition was performed both in the "classical" way, without using word-boundary information and rescoring, and with the joined word-boundary detector module and rescoring. With the latter one, the ratio of correctly recognized words was higher by approximately 3.8% when compared to the baseline system. In general, the recognition rate for each report (of the 20 reports used) increased in accuracy, however, recognition performance with the extended ASR system decreased for two reports. A deeper error analysis revealed that this was due to improper prosodic segmentation related to errors made by the fundamental frequency extraction tool (pitch tracker).

3.3 Sentence type recognition and clause boundary detection

In classical statistical speech recognition, the identification of the sentence type or modality is not possible. To determine them, syntactic and semantic analyser tools can be used (see Shriberg et al. 2000). These systems also use ASR output, so all recognition errors are passed into the modality classifier module. Moreover, it is possible that prosody is the solitary feature identifying the sentence type (Kompe 1997). In written language, punctuation marks help to identify the sentence type. However, commas do not always coincide with the prosodically marked grouping of clauses (Olaszy 2005).

Sentence type recognition is also a semantic level task, which is based on the classification of the intonation of the sentence. Its syntactic requirements are sentence and/or clause level speech segmentation. After this, the sentence type can be recognized.

The sentence type models for Hungarian used were: declarative (S), explicit question (K), yes/no question (E), imperative/exclamatory (X), optative (O), andfor German: declarative (S), interrogative (E), imperative/exclamatory (X). As their intonation is different, yes/no questions and questions requiring an explicit answer were separated in Hungarian. Exclamatory and imperative sentences are merged in both languages, as they were not found to be significantly different in intonation. This merging did not influence the corresponding punctuation mark (i.e., an exclamation mark was used for both sentence types). Non-terminal clauses (T) and silence (U) were also modelled to allow an alignment for speech parts containing several clauses and sentences.

The iteratively optimized sentence type recognizer yielded the results for Hungarian (in Figure 2) and German (in Figure 3). Correctness was evaluated for each sentence type separately.

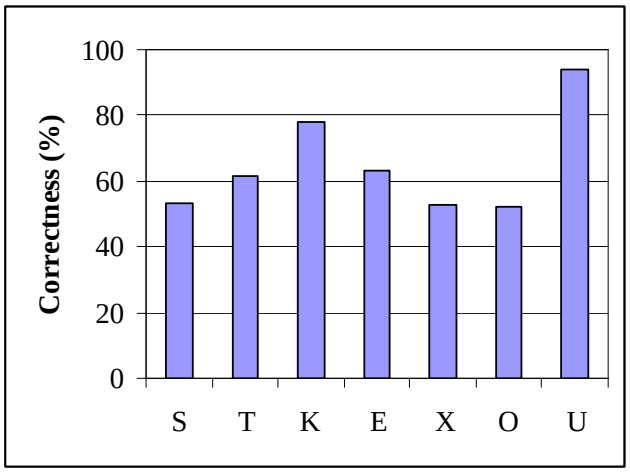


Figure 2. Correctness (in %) evaluated separately with 11 state HMMs using a time window of 40 frames (400 ms) on the Hungarian Children Database

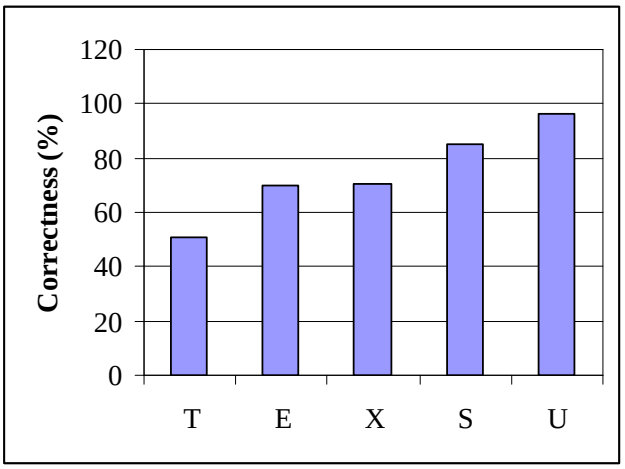


Figure 3. Correctness (in %) evaluated separately with 11 state HMMs using a time window of 40 frames (400 ms) on the Kiel Corpus (German)

4 Discussion and conclusions

The article addressed the role of prosody in automatic speech recognition, with a special attention paid to fixed-stress languages. A novel prosodic segmentation method was briefly overviewed, which is based on alignment of phonological phrases. Since the boundaries of the phonological phrases coincided with word-boundaries in the languages tested, partial word-boundary detection could be

performed. This process can help to improve the performance of automatic speech recognizers (ASR). The segmentation into phonological phrases is believed to help automatic syntactic and semantic processing of human speech (Cristophe 2004). HMM-based phonological phrase alignment was implemented for Hungarian and Finnish (both fixed-stress languages). The performance of the phonological phrase alignment was analysed using precision and recall rates for word-boundary detection, although not all word boundaries were phonological phrase boundaries. The reason for speaking about word-boundaries is that in current ASR applications, phonological phrase boundaries cannot be used, only word-boundaries are taken into consideration. However, prosodic segmentation improved the performance of a Hungarian ASR application by 3.8%. Like phonological phrase alignment, larger units, such as clauses, can be also aligned based on prosodic modelling. After this alignment, clauses can be classified as sentence-terminal and non-terminal clauses, and the type of the sentence can be recognized. This facility can contribute to automatic semantic analysis, to the automatic placement of punctuation marks in ASR systems, or to the classification of sentence type (for example, interrogative vs. affirmative, which is very important in dialogue systems). A similar framework can be also used for automatic emotion recognition in speech (Tóth et al. 2008) – which is based on acoustic analysis of supra-segmental prosodic features, which have been expanded by some segmental features.

References

Cristophe, A., Peperkamp, S., Pallier, C., Block, E. & Mehler, J. 2004. Phonological phrase boundaries constrain lexical access I. Adult data. *Journal of Memory and Language*, 51: 523-547.

Hirose, K., Minematsu, N., Hashimoto, Y. & Iwano, K. 2001. Continuous speech recognition of Japanese using prosodic word boundaries detected by mora transition modeling of fundamental frequency contours. In *Proceedings of ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding.* Red Bank, NJ, USA. 61-66.

Hunyadi, L. 2002. *Hungarian sentence prosody and Universal Grammar*. New York: Peter Lang.

Jelinek, F. 1998. *Statistical methods of speech recognition*. MIT-Press, USA.

KIEL Corpus of read speech, Volume I. 1994. Institut für Phonetik und digitale Sprachverarbeitung der Christian-Albrechts-Universitat zu Kiel.

Kompe, R. 1997. *Prosody in speech understanding systems*. LNAI 1307, Springer.

Kompe, R., Kießling, A., Niemann, H., Nöth, E., Schukat-Talamazzini, E. G., Zottmann, A. & Batliner, A. 1995. *Prosodic scoring of word hypotheses graphs.* In *Proceedings of European Conference on Speech Communication and Technology*. Vol. 2. 1333-1336.

Mandal, S., Gupta, B. & Datta, K. 2007. *Word boundary detection based on suprasegmental feaures, a case study on Bangla speech*. *International Journal of Speech Technology*, 9(1-2): 17-28.

Olaszy, G. 2005. Prozódiai szerkezetek jellemzése a hírfelolvasásban, a mesemondásban, a novella- és a reklámok felolvasásában. (= Characterization of prosodic structures in the reading of news, tales, short stories and advertisements). *Beszédkutatás 2005* (= Speech Research 2005), 21-50.

Shriberg, E., Stolcke, A., Hakkani-Tür, D. & Tür, G. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2): 127-154.

Szaszák, Gy. 2009. A szupraszegmentális jellemzők szerepe és felhasználása a gépi beszédfelismerésben (= The role and usage of supra-segmental features in automatic speech recognition). PhD dissertation. Budapest, Hungary: Budapest University of Technology and Economics.

Tóth, Sz. L., Sztahó, D. & Vicsi, K. 2008. Speech emotion perception by human and machine. In: *Proceedings of International Conference of COST Action 2102, Patras, Greece, October 29-31, 2007*. Springer. 213-224.

Vainio, M., Altosaar, T., Karjalainen, M., Aulanko, R. & Werner, S. 1999. Neural network models for Finnish prosody. *Proceedings of ICPhS 1999*. San Francisco. 2347-2350.
Varga, L. 2000. A magyar mellékhangsúly fonológiai státusáról. (= On the phonological status of secondary stress in Hungarian). *Magyar Nyelvőr*, 124(1): 91-108.
Veilleux, N. M. & Ostendorf, M. 1993. Prosody/parse scoring and its application in ATIS. In *Proceedings of ARPA Human Language Technology Workshop '93*. 335-340.
Vicsi, K. & Szaszák, Gy. 2005. Automatic segmentation of continuous speech on word level based on supra-segmental features. *International Journal of Speech Technology*, 8(4): 363-370.
Vicsi, K. & Vig, A. 1998. Az első magyar nyelvű beszédadatbázis (= The first Hungarian speech database). *Beszédkutatás '98* (= Speech Research '98), 163-177.
Waibel, A. 1988. *Prosody and speech recognition*. London, UK: Pitman.

REPORT ON BUILDING A TOOL FOR ROMANIAN SPONTANEOUS SPEECH RECOGNITION

1 Introduction

The personalized interaction between human subjects and computers presents a significant challenge today, as software services and products become more and more user-centered. Thus, spoken computer dialogue constitutes one of the most natural and convenient means of interaction for the human.

For several international languages (i.e., English, French), there are complete human-computer dialogue systems, in domains from train or plane ticket reservation (the American system CMU Communicator, designed at Carnegie Mellon University in the last decade of the 20th century, systems developed by France, in the ESPRIT European projects, in the last two decades of the 20th century), to resource –meeting room management in voice portal applications (the PVE – „Portail Vocal pour l'Entreprise" system, developed in cooperation with Grenoble University 1, CNRS and several companies, such as France Telecom, and with government financing, in 1992-2005). On the other hand, in other languages such as Romanian, considered "under-resourced, from a speech database point of view" (according to recent studies), the development of spoken dialogue systems is a long-term process.

The task of the speech recognition component in a spoken dialogue system consists of converting the utterance (in acoustic form) from the user into a sequence of discrete units, such as phonemes (sound units) or words. A major obstacle in accomplishing a reliable recognition is speech signal variability, which results from the following factors:

• Linguistic variability, in which includes the effects of several linguistic phenomena, such as phonetic co-articulation (i.e., the fact that the same phoneme can have different acoustic realizations in different contexts);

• Speaker variability, in which includes the effects of inter- and intra-speaker acoustic differences; inter-speaker differences are determined by physical factors, such as the particular shape of the vocal tract, the age, sex or origin of the human subjects (the fact that a speaker may not be a native speaker of the language being used for communication). Intra-speaker differences occur when same word can be uttered in several ways by the same speaker, according to her or his emotional or physical state, or to the pragmatic (and situational) context of the utterance – for instance, a word can be uttered more emphatically in order to stress a certain idea;

• Channel variability: includes the effects of environmental noise (which can be either constant or transient) and of the transmission channel (e.g., microphones, telephone lines, or data channels – "Voice over IP").

The speech recognition component in a typical dialogue application has to take into account several additional issues:

• Speaker independence: since the application is normally used by a wide variety of individuals, the recognition module may not be trained for a single speaker (or for a few speakers) supposed to use the system, as in voice dictation applications. Thus, speech has to be collected from an acoustically representative set of speakers, and the system will use these data in order to recognize an utterance from (potential) users, whose voices were not used during training. This is why the performance of the speaker independent recognition process is generally poorer than that for speaker dependent recognition.

• Size of the vocabulary: the number of words that are "intelligible" to the dialogue system depends upon the application considered, as well as on the complexity of the dialogue. Thus, a strictly controlled and rather inflexible dialogue may constrain the user to a small vocabulary, limiting the system to a few words expressing the available options . Yet, in more natural and flexible dialogues, the accessed vocabulary can process several thousand words (for instance, the PVE system, developed in France as a voice portal for enterprises, has about 6000 words recognition module).

• Continuous speech: the users are expected to be able to establish a conversation with the spoken dialogue system, using unconstrained speech and not commands uttered in isolation (isolated voice command systems for industrial robots have been developed also in Romania, in the 1980s, at University "Politehnica" of Bucharest). The issue of establishing the boundaries of words is extremely difficult in continuous speech, since in the acoustic signal, there is no physical border between words. Hence, linguistic or semantic information has to be used in order to separate the words.

• Spontaneous speech: since users' utterances are normally spontaneous and unplanned, there are generally characterized by disfluencies, such as hesitations or interjections (e.g., "hmm"), false starts, in which the speaker begins an utterance, stops in the middle and re-starts, or extra linguistic phenomena, such as a cough. The speech recognition module must be able to extract a word sequence out of the speech signal thereby allowing the semantic analyzer to deduce the meaning of the user's utterance.

The concern of developing a speech recognition module suited for spoken dialogue systems is not trivial, although stand-alone continuous speech recognition systems exist (e.g., for dictation applications) and are beginning to appear in Romania. Thus, a continuous speech recognition module for spoken dialogue in Romanian must be adapted to the spontaneous nature of speech, and be able to interact with other modules in the dialogue systems, namely the semantic analyzer (which can decide whether the utterance is meaningful or not) and the dialogue manager (which can decide whether the utterance is reasonable and relevant to the dialogue history).

2 Speech recognition systems

In principle, speech recognition involves finding a word sequence, using a set of determined models acquired in a prior training phase and matching those models to the input speech signal. For small vocabularies (i.e., less than 100 words), these models can capture word properties, but tend to focus on sound units are generally modeled (such as phonemes or triphones, which represent phonemes in the context of neighboring phonemes). The most successful approaches consider this process to be probabilistic and has to account for temporal variability (due to different sound durations) and acoustic variability (due to linguistic, subjective and channel-related factors, as emphasized above). Such systems, based on statistical approaches, are available in the research community (the SPHINX system from Carnegie Mellon University, the HTK – "Hidden Markov Modeling Toolkit" toolkit from Cambridge University, the RAPHAEL system, from Laboratoire d'Informatique de Grenoble, etc.), as well as commercially (systems developed by Nuance, Dragon or Microsoft in the United States; those developed by France Telecom, Prosodie in France etc.). At the same time, the first continuous speech recognition system for the Romanian language was developed in 2006 at the Military Technical Academy in Bucharest. Research to improve some of the components in this system has been conducted at University "Politehnica" of Bucharest.

A typical continuous speech recognition system works in two areas:

• training, which involves the creation of the necessary acoustic and linguistic knowledge, for the models being used,

• recognition, which involves the resources created during training which convert a spoken utterance (in its acoustic form) into a word sequence.

For practical reasons, we have chosen to use "Hidden Markov Modeling Toolkit" (HTK) as the basis for a general speech recognition front-end architecture in a spoken dialogue system. Before explaining the details of the architecture, we start with a brief statement of the constraints that this type of architecture should satisfy:

• the nature of the speech signal: spontaneous speech, in the Romanian language;

• system's usage conditions: in closed laboratory room, with a constant signal-to-noise ratio (SNR > 25 dB);

• microphone type: headset, with constant distance between speaker's vocal cavity and the microphone;

• vocabulary size: between 3000 and 10000 domain-independent words;

• characteristics relative to the speaker: speaker-independent, but tuneable to the voice of a particular speaker, with less than 60 seconds of speech signal, acquired from this particular speaker;

• voice detection: "on-line" speech decoding, with automatic "voice activity detection" (VAD); hence, "push-to-talk" setups will not be used.

In this context, we have designed a sequential architecture, based on HTK, with the following specific features:

a) speech signal parameterization (for both training and decoding regimes): MFCC coefficients;

b) acoustic modeling: triphone-level; for each triphone, an HMM is trained;

c) HMM features for each Romanian triphone:

• number of states: 5 (including one initial state and one final state, both non-emitting, and three emitting intermediary states);

• HMM topology: left-right, where transitions towards remote states and "backwards" transitions are not allowed;

• output distribution: continuous, as a weighted linear combination of Gaussian mixtures, for each emitting state.

Thus, the system is designed to work in two ways, training and testing (recognition).

3 Romanian speech databases

3.1 Basic principles

A database for continuous speech recognition is made up of the following components:

• a set of speech signal samples;

• a set of correspondences between the speech signal samples with their features (i.e., duration of the signal, identities of the speakers, speech type – read, spontaneous, etc.);

• a set of labels that provide the words or phonemes that are uttered in each speech segment;

• a set of acoustic parameters, which "synthetically" represent the speech signal (Mel Frequency Cepstrum Coefficients – MFCC, Perceptual Linear Prediction Coefficients – PLP, etc.); a set of acoustic parameters is associated to each (suitably chosen) speech signal window.

A phonetic dictionary is not necessarily part of a speech database, but such a dictionary (that translates between the words and its component phonemes) is always needed when building a continuous speech recognizer. Linguists' research showed that 36 phonemes are enough to cover the various word pronunciations in the Romanian language. In one of our collaborations with expert linguists, we've obtained an extensive phonetic dictionary that contains a total of 800,000 entries (literary word forms). This dictionary does not cover all the various word pronunciations encountered in the spoken (oral) language, but it gives us a head start and the rest of the work focuses on matching the pronunciations with the word forms encountered during the labeling process.

Although we are using IPA notations to distinguish between different phonemes, we've decided to employ other phoneme notations for our "in-house" development and testing. We determined that most of the tools work easier with ASCII encoded text files versus Unicode encoded files. Table 1 summarizes the correspondence between IPA and our SpeeD phoneme notations.

A database can be acquired via the following methods:

1. direct recording; this yields a series of specific issues:

• choosing the recording place (studio, laboratory, etc.);

Table 5. Corresponding between IPA and our phoneme notations

IPA Symbol	SpeeD Symbol		IPA Symbol	SpeeD Symbol		IPA Symbol	SpeeD Symbol
a	a		o̞	o1		f	f
ə	@		W	w		v	v
e	e		C	k2		h	h
i	i		ɓ	b		ʒ	j
ʲ	i1		p	p		ʃ	s1
ɨ	i2		k	k		l	l
o	o		t͡ʃ	k1		m	m
u	u		g	g		n	n
y	y		d͡ʒ	g1		s	s
ø	o2		ɟ	g2		z	z
e̞	e1		d	d		r	r
j	i3		t	t		t͡s	t1

• choosing the microphone (microphone type, signal to noise ratio, noise filtering);

• choosing the recording workstation (sound acquisition board, signal amplitude, overall signal to noise ratio, etc);

2. labeling audio books or other spoken materials; the specific issues in this situation are:

• leveling the differences in the sampling frequencies for the different spoken materials;

• splitting the audio and labeled content into small speech entities: phonemes, words or word groups.

• detecting and correcting labeling and basic speech entities splitting errors;

3. recording Internet broadcasted TV or radio shows + labeling the content; the specific issues in this situation are:

• homogeneity of the recording conditions (outdoor shows, studio recordings, movies, etc.);

• leveling the differences in speech coding standards (A - PCM, μ- PCM, etc.);

• leveling the differences in the sampling frequencies for the different spoken materials;

• controlling the speaker set (so that a balanced amount of speech is obtained from all speakers);

• detecting and correcting labeling errors;

4. direct acquisition from radio or TV broadcast channels + labeling the content; in this case, the specific issues are:

• the analog-digital conversion of the signal;

• the homogeneity of the recording conditions;

• detecting and correcting labeling errors.

Our team has focused its efforts in acquiring speech databases using the first three of the previously mentioned methods.

3.2 Databases descriptions

This section describes the various databases we've acquired as a starting point for our speech recognition projects. Here are some notes about the purpose and the features of each database:

• Database 1 and 2 were acquired using methods 2 and 3 and are most suitable for tuning up a system for continuous and spontaneous speech recognition;

• Database 3 is the starting point for any projects that aim to model phonemes or triphones. We've acquired this database in order to address the initialization of the system.

• Database 4 has been built using the first acquisition method (direct recording) in order to offer an enhanced control of the words being uttered, the speakers, recording environment, etc.

• Database 5 is a small database we've acquired for a different project: a voice driven remote controlled device that understands and executes only a few basic voice commands.

3.2.1 Database 1 – Romanian language spontaneous speech recognition project, see Table 2.

Table 2. Database 1.

Acquisition method	Recording Internet broadcasted TV or radio shows + labeling the content (3)			
Acquisition date	2008			
Authors	Andi Buzo, Cristina Petrea, Diana Hanes, Florin Baltescu, Roxana Faur			
Speech signal	Language		Spoken Romanian	
	Type		Oral, continuous, spontaneous	
	Total duration		Approximately 4 hours	
	Recording environment		TV studio	
	Sampling frequency		8 kHz	
	Sample size		16 bits	
	Labeling		Word groups level (60 seconds)	
Speakers	Number of speakers	12	Females	8
			Males	4
	Sessions per speaker		3-20	
	Time between recording sessions		One day to two weeks	
Words	Total occurrences		37604	
	Number of different words		8068	

3.2.2 Database 2 – Romanian language spontaneous speech recognition project, see Table 3.

Table 3. Database 2.

Acquisition method	Labeling audio books or other spoken materials (2)			
Acquisition date	Summer 2009			
Authors	Adina Popa, Diana Uzum, Mihai Iordache, Horia Cucu, Dan Oneata, Tudor Mihailescu, Ioana Rolea			
Speech signal	Language		Literary Romanian	
	Type		Read, continuous	
	Total duration		Approximately 11 hours	
	Recording environment		Recording studio	
	Sampling frequency		16 kHz	
	Sample size		16 bits	
	Labeling	3% word level 12% word groups level (up to 3 seconds) 85% word groups level (60 seconds)		
Speakers	Number of speakers	7	Females	3
			Males	4
	Sessions per speaker		Unknown	
	Time between recording sessions		Unknown	
Words	Total occurrences		40016	
	Number of different words		7770	

3.2.3 Database 3 – Romanian language spontaneous speech recognition project, see Table 4.

Table 4. Database 3.

Acquisition method	Labeling audio books or other spoken materials (2)	
Acquisition date	Autumn 2009	
Authors	Adina Popa, Diana Uzum, Mihai Iordache, Tudor Mihailescu, Ioana Rolea, Florin Baltescu	
Speech signal	Language	Literary Romanian
	Type	Read, continuous
	Total duration	N/A
	Recording environment	Recording studio
	Sampling frequency	16 kHz
	Sample size	16 bits
	Labeling	phoneme level

Speakers	Number of speakers	10	Females	3
			Males	7
	Sessions per speaker		Unknown	
	Time between recording sessions		Unknown	
Words	Total occurrences		N/A	
	Unique occurrences		N/A	

3.2.4 Database 4 – Romanian language spontaneous speech recognition project, see Table 5.

Table 5. Database 4.

Acquisition method	Direct recording (1)			
Acquisition date	Spring 2010			
Authors	Adina Popa, Diana Uzum, Tudor Mihailescu, Ioana Rolea, Florin Baltescu			
Speech signal	Language		Spoken Romanian	
	Type		Read, single words only	
	Total duration		N/A	
	Recording environment		Laboratory	
	Sampling frequency		16 kHz	
	Sample size		16 bits	
	Labeling		word level	
Speakers	Number of speakers	5	Females	3
			Males	2
	Sessions per speaker		10 -20	
	Time between recording sessions		Couple of hours to couple of days	
Words	Total number of words		50000	
	Number of different words		10000	

3.2.5 Database 5 – voice driven remote control device project, see Table 6.

Table 6. Database 5.

Acquisition method	Direct recording (1)
Acquisition date	Spring 2010
Authors	Andreia Vlad, Florin Teodoru, Daria Ion, Daniela Milea

Speech signal	Language		Spoken Romanian	
	Type		Read, single words	
	Total duration		N/A	
	Recording environment		Laboratory	
	Sampling frequency		16 kHz	
	Sample size		16 bits	
	Labeling		word level	
Speakers	Number of speakers	4	Females	3
			Males	1
	Sessions per speaker		3 -10	
	Time between recording sessions		Couple of hours to couple of days	
Words	Total number of words		5600	
	Number of different words		4	

4. Training strategy

Speech recognition using statistical models is basically a decision-making process. At the end of the process, the following question is answered: “Which model or sequence of models from a given set better match with a given speech signal?”. Statistical models are mathematical models that reflect various speech features. One such models is the Hidden Markov Model (HMM) that we have been using in our projects. The process of building the models is called training. Several speech elements can be modeled, like phonemes, triphones (a phoneme for which the leading and following phones are specified), syllabuses or even words. The selection of speech elements to be modeled is an important decision in the training process as recognition performance will vary significantly. After a series of experiments we have decided that the best way to build speech models is to follow these steps:

1. Phoneme models are trained using the isolated phoneme (phoneme labeled) database.
2. Phoneme models resulted from step 1 are further trained using the isolated words database.
3. Triphone models are built by cloning the model of the central phoneme of the triphone.
4. Triphone models are trained by using the isolated words (word labeled) database.
5. Triphone models are trained by using both word-labeled and the file labeled database (used for continuous speech recognition).

In order to train the models properly, a correspondence between the models and the recorded speech files must be established. This correspondence is realized by labeling the boundaries between phonemes in a word. (Phonemes are usually uttered

inside words and not separately, so the best way to train phoneme models is to use utterances within words and not isolated, spoken phonemes). Labeling each phoneme in each word is a time-consuming process and collecting a large database becomes very difficult. In order to overcome this issue, **embedded training** is used which applies the Baum-Welch algorithm. This algorithm is able to align models to the recorded data. However, the use of well-trained models instead of plain (untrained) models in initialization helps the algorithm to achieve a better alignment and obtain a faster convergence. That is why step 1 is invoked.

Triphones are used because, by the way they are defined, they add more restrictions during the recognition process. Once it has been decided that a model is most probable, practically the decision about its neighbors is affected. The triphone model specifies not only a certain phoneme but also the phonemes that surround it. The first triphone models are cloned from the phoneme models. While training, the central states of the HMMs for triphones with the same central phoneme are tied, which means that at the end of the training, the central states will have the same parameters.

Another important topic to be mentioned is why steps 4 and 5 are not merged. There are differences among words used for isolated word recognition and continuous speech recognition. Words in a sentence are pronounced differently depending on the syntactical role of the word in the sentence or depending on the type of sentence (affirmative, interrogative, imperative, etc.). All these changes affect the models characteristics and should be taken into consideration. If we skipped step 4, models would not have been well aligned with the data because the models are not very robust at this stage.

5 Recognition

Generally, the recognition process considers input to be an unlabeled segment of speech signal (because it is unknown to the system) and it involves three distinct stages:

1. Acoustic processing of the input speech signal, in a manner identical to the training process.

2. Acoustic decoding of the parameter vectors sequence representing the input signal. This process matches the acoustic vectors to the models estimated during training and obtains a sequence of the most likely acoustic units associated with the original utterance.

3. Refining the results of the acoustic decoding process, using the language model. This step matches the word sequence obtained in the preceding step, to the language model. Thus, if the word sequence was found using only acoustic models and it is highly unlikely according to the language model, then the system would choose less likely word sequences (according to the acoustic decoding) that are more likely, according to the language model.

At the end of these processing steps, the output of the recognition system is represented by a number of alternative word sequences for an utterance. Sometimes,

the differences between alternative word sequences are small and determined by semantically irrelevant words (for example, two alternatives for the same original Romanian utterance, *Când decolează avionul?* – 'when does the plane take off?', may be *la cât decolează avionul* – 'at what time does the plane take off', or *pe când decolează avionul* 'around what time does the plain take off', but also *când deraiază vagonul* – 'when does the coach derail', which obviously has a different meaning than the original utterance). The choice of the most relevant alternative, in a specified context, is the responsibility of other components of the dialogue system (namely, the semantic analysis component, or the dialogue manager). For example, if the application domain is air traffic, then the last version, indicated above, will be eliminated as irrelevant.

In the recognition stage, the resources created during training are used. More precisely, the HMMs and the phonetic dictionary are used for recognizing the utterances. The process encompasses the following steps:

1. Manual definition of a set of test sentences.
2. Production, acquisition and parameterization of utterances of test sentences.
3. Triphone-level decoding, which consists of determining the set of HMMs that have generated (with maximal probability) the sequence of acoustic parameters obtained at step 2. This process involves a Viterbi alignment of the sequence of acoustic frames to the set of paths in the HMMs, and then chooses those for which the Viterbi distance is minimal (i.e. the sequence of acoustic vectors had been generated by the considered HMMs with maximal probability). This is accomplished with the "HVite" HTK tool.
4. Converting the set of triphones obtained at step 3 in a word sequence, according to the phonetic dictionary (which, this time, is used in a different way than for training: now, triphone sequences are converted to words). This is also done with the "HVite" tool.
5. Performance evaluation: this process involves iterating stages 1 to 4 on a sufficiently large set of test utterances, and then comparing, by dynamic alignment (the Levenshtein distance), the reference sentences with the recognition output. Thus, we compute both an utterance error rate and a word error rate. This is done with the "HResults" HTK tool.

For improving the efficiency of the training process (with respect to the size of the database), state tying might be used, or at least mixture tying can be envisaged for the HMMs, in order to reduce the number of parameters to be estimated.

6 Results

Several outputs may result from the processes described above. Some examples are presented below.

6.1 Triphone occurrence statistics

As an example, the database described in Section 3.2.1 (Database 1) is meant to be used in an application for spontaneous speech recognition purposes. The results

that characterize the database are represented by statistics referring to the number of occurrences for both words and triphones.

One result that may be found during database modeling is a histogram representing the number of occurrences for different triphones in the database. An example is shown in Figure 1.

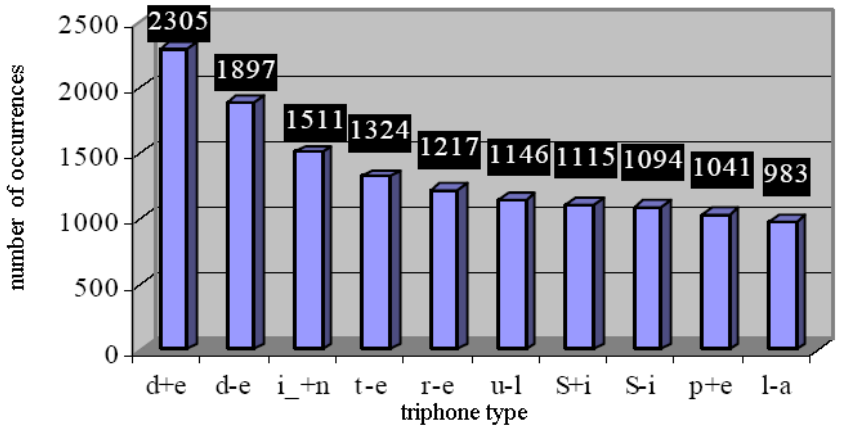


Figure 1. The triphones with the highest number of occurrences in Database 1.

6.2 Word occurrence statistics

In order to validate the speech recognition architecture, several experiments were conducted with a reduced word set, as when building a recognition system for a language with a very small vocabulary. The speech sequences were chosen from the same speaker, also based on Database 1. The seven Romanian language words chosen for this experiment are listed in Table 7.

Table 7. Limited set of (frequent) Romanian words chosen in the recognition experiments

Words	**Occurrences in the first phase**	**Occurrences in the second phase**
în	27	34
şi	22	25
de	36	51
la	24	32
cu	12	21
din	14	22
un	10	20

6.3 Recognition rate

For Database 1, the training of the HMMs was performed in two different phases, shown in Table 7. The second phase utilizes more observation data, about 40% more

than the first phase. The goal here was to show how a greater number of word occurrences yields a higher word recognition rate. The recognition rate used as metric in this experiment was calculated as the number of words correctly recognized over the total number of words present in the test sequences.

In these conditions the recognition rate for the first phase was 52%, while the recognition rate for the second phase was 70%. It has been observed that the order in which the triphones are trained slightly affected the results, so these figures represented the best match.

For the larger databases (2–4), the recognition process has been slightly improved (see Table 8). The recognition process outputs several words that have the highest probabilities to be the same as the sample to be recognized. Then, the comparison is made and recognition rates are calculated taking into account all the words selected as possible "candidates".

Table 8. Example of recognition rates on Database 4.

Training/ testing iteration	Fully recognized [%]	Partially recognized [%]	Partially recognized among first two results [%]	Partially recognized among first three results [%]	Partially recognized among first four results [%]
6	39.03	60.76	63.28	65.79	66.7
11	48.89	67.61	70.62	72.33	74.25
16	45.67	69.11	71.83	74.14	74.85
21	44.87	70.12	72.64	74.55	75.45
26	44.87	70.02	73.64	75.25	76.06
31	44.97	70.52	73.54	75.65	76.46
41	45.88	70.93	73.04	75.25	76.36
51	46.38	70.62	73.44	75.05	75.96
61	46.08	70.12	73.44	75.25	76.16
71	46.08	70.12	73.34	74.95	75.86
81	46.38	70.52	73.14	75.15	75.86
91	46.28	70.62	73.24	75.05	76.16

7 Conclusions and prospects

A medium-size database for the Romanian language was created by gathering together all our SpeeD databases. In order to improve the analytic possibilities, the goal is to increase the size of the database. Also, we are working on improving the recognition process. The "measurable" result we are looking for is a higher recognition rate.

Some conclusions drawn from the statistical means from the first corpus: the large number of triphones that have been identified in Database 1 is specific to spontaneous speech. There are cases when some triphone combinations have a greater number of occurrences, compared to others. For instance, triphones "d+e", "d-e" and "i_+n" have more than 1500 occurrences. This has a logical explanation,

as the most frequent triphones happen to be prefixes and suffixes. The frequency of these triphones could also be the result of natural hesitations in speech planning. For instance, when stammering, the speaker tends to double the beginning or ending of a word. Likewise, a person who is speaking in front of an audience or an important person, simply being a nervous person, or in a situation that involves a higher level of stress or even attention, tends to hesitate. In this situation, in order to regain speech fluency and avoid the conversational blocking, the speaker tends to double some of the verbal constructions, in an unconscious manner. While the speaker may double, triple or multiply produce a whole word or phrase, he gains precious time that allows him to construct his utterance.

In a similar way, when tension rises in a monologue or a conversation, a normal healthy person may adopt more verbally erratic behavior. In this case, the physiology of the person changes, the face and body contract and the person tends to sweat, to become red or they may become pale. This anxiety could cause different behaviors, but the most predominant one is disfluency. The most common pattern here is the repeated pronunciation of the first syllables of the words. Hence, spontaneous speech may be a way to observe common speech patterns in persons with speaking anxiety.

The histograms determined for the word occurrences in the database revealed that: “de” (1659), “la” (891), “in” (803), “a” (667), “şi” (656) were the most frequent. This observation, together with having a closer look at some triphone constructions and their number of occurrences leads to some other conclusions:

• “d+e” (2305) and “d-e” (1897) are part of the word “de”, which is a preposition in the Romanian language. The high number of occurrences of both “d+e” and “d-e” demonstrates that “de” is the preposition mostly used in the recorded Romanian database. Subtracting the number of occurrences for the word “de” from the triphones’ frequencies: 2305 – 1659 = 646 and 1897 – 1659 = 238 mean that there are more words ending with “d+e” than words beginning with “d-e”.

• “i_+n” (1511) and “i_-n” (852) also form the Romanian preposition “în” and, following the same reasoning as above, the differences 1511 – 803 = 708 and 852 – 803 = 49 revealed the number of occurrences of other words besides “în”, ending with “i_+n” and beginning with “i_-n”. There is a higher number of words ending with “i_+n” than words beginning with “i_-n”.

• “S+i” (1115) and “S-i” (1094) form the conjunction “şi” in the Romanian language. The differences 1115 - 656 = 459 and 1094 – 656 = 438 represent the number of occurrences for other words besides “şi”, where the words ending with “S+i” have more occurrences than the ones beginning with “S-i”.

• “l+a” (982) and “l-a” (983) have almost the same number of occurrences, most of them (891) belonging to “la” preposition. In this situation, almost the same number of words from the database begin or end with “l+a” / ”l-a”.

There are 2903 triphones with less than 10 occurrences and they represent 57% of the total number of triphones. When enlarging the size of the database, it is expected that the triphones with lower frequencies will have more entries. Due to the nature

of spontaneous speech, it is expected to obtain a completely different view of triphones occurrences after augmenting the database.

The corpus and the spontaneous speech recognition results have applicability as recognition tools. They will ease the work of the user, taking off some of the existent constraints, like forcing a person to use grammatically correct phrases or by stressing an individual with dyslexia to correctly select his words.

In future studies, the first problem considered will be database augmentation: nowadays, hundreds of hours are spent at the acoustic level, in state-of-the-art systems.

Secondly, the tradeoff between recognition accuracy and computing (decoding) time has to be carefully considered. In dialogue applications, the former can be sacrificed for the latter, to a greater extent than in the case of related applications as in broadcast news transcription. Hence, the computing time can be improved (reduced) with parallelization strategies, by extending previous work, or by adopting particular search heuristics while decoding the signal. Database 4 is currently used for recognition purposes and for time optimizations. These experiments are currently in progress.

Lastly, it might also be interesting to improve the robustness of the system at the acoustic level, by computing confidence measures on the phoneme or triphone-level HMM probability estimates. This work is currently in progress.

Prof. Corneliu Burileanu,
Cristina-Sorina Petrea (PhD student), Andi Buzo (PhD student),
Horia Cucu (PhD student), Alina Pasca (PhD student)
Faculty of Electronics, Telecommunications and Information Technology,
University “Politehnica” of Bucharest, Romania
e-mail: corneliu.burileanu@yahoo.com

TOOLS USING SPEECH TECHNOLOGY FOR RESEARCH AND EDUCATION

Glottalizer: A tool to transform regular voice into irregular voice

This paper describes a freely available[1] software program that semi-automatically transforms regularly phonated speech into an irregular voice. This program utilizes a transformation method that introduces irregular pitch periods into a modal speech signal by scaling the amplitude of the individual cycles. The scaling factors can be set individually or 'copied' as a pattern from irregularly phonated speech.

1 Irregular phonation

Irregular phonation refers to regions of voicing where there are significant periods of abrupt, cycle-to-cycle changes in either the spacing of the glottal impulses and/or their amplitudes (see Figure 1c). This type of irregular vocal fold vibration includes deviations from periodicity which exceed the usual jitter and shimmer values present in regular or modal phonation (Surana & Slifka 2006). This deviation is clearly audible to people with normal hearing. Irregular phonation also includes when the fundamental frequency abruptly drops below the speaker's characteristic voice register, resulting in a perceivable change in voice quality. Irregular phonation can serve as a cue to segmental contrasts (Laver 1994, p. 330-331) and prosodic structure (Dilley et al. 1996), as well as to affective state (Gobl & Ní Chasaide 2003) and speaker identity (Henton & Bladon 1987).

2 Transformation method

The transformation method utilized to produce irregular voice introduces irregular pitch periods into a modal speech signal by scaling the amplitude of individual cycles within the speech signal. First, the periods are separated by windowing, then multiplied by appropriately chosen scaling factors, and finally overlapped and added (the details of the transformation method are described later in this section). Thus, amplitude irregularities are introduced via boosting or attenuating selected cycles. The abrupt, significant changes in cycle lengths that are characteristic of naturally-occurring irregular phonation can be achieved by removing (scaling to zero) one or more consecutive periods. A method is proposed to copy 'stylized pulse patterns' (the spacing of glottal pulses and their amplitudes) in order to set the scaling factors semi-automatically.

The input to the transformation method is the speech waveform ($x[n]$) with markers for the approximate times of glottal excitations (*pitch marks*, denoted by p_i, $1 \leq i \leq P$) in the region to be transformed. As a first step, a rough approximation of the impulse response for each glottal pulse is extracted by applying an asymmetric Hanning-window ($w_i[n]$) in the vicinity of each pitch mark. The peak of the window

1. http://www.bohm.hu/glottalizer.html

is positioned on the pitch mark (p_i) and it spans from the previous (p_{i-1}) to the next pitch mark (p_{i+1}). This windowing procedure is the same as the analysis stage in the PSOLA algorithm (Moulines & Charpentier 1990). It then extracts rough estimates of the individual glottal cycles into separate waveforms. The samples in each of these one-impulse-response waveforms are then multiplied by a hand-selected scaling factor (s_i) and overlapped-and-added to re-synthesize the signal (Figure 2):

$$\hat{x}[n] = \sum_{i=1}^{P} \left(s_i x[n] w_i[n] \right)$$

where $\hat{x}[n]$ is the output speech signal. The term 'overlap-and-add' refers to the way the scaled one-impulse-response waveforms are recombined into one long waveform: each of the scaled signals is positioned so that the timing of the pitch marks is the same as in the input. As a result of this positioning, the waveforms overlap in time. The final output is calculated by adding the samples that fall into the same instant in time.

The scaling factors can either boost ($s_i > 1$), attenuate ($s_i < 1$), remove ($s_i = 0$) or leave the impulse responses unmodified ($s_i = 1$). In regions of the speech waveform where all scaling factors are set to one, the original signal is reconstructed (apart from rounding errors), so any possible artefacts are limited to the amplitude-manipulated regions of the speech signal. See Figure 1b for an example of a transformed speech waveform.

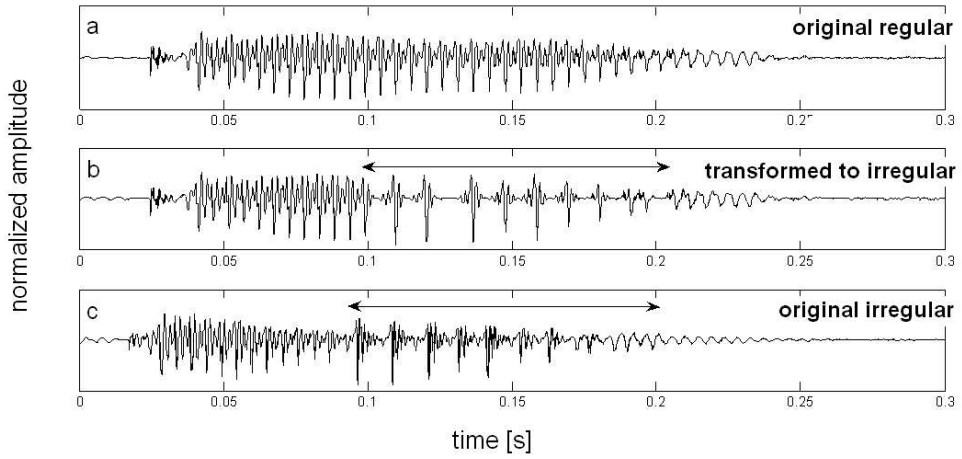


Figure 1. A speech recording with a regular ending (a) and its transformed version (b). An originally-irregular recording is also shown (c). Horizontal arrows mark irregular regions.

Note that the method does not alter the timing of the glottal pulses (which is what the PSOLA algorithm does in order to change F0). In contrast to PSOLA, where the

aim is to implement fine adjustments of the fundamental frequency, here we need abrupt, substantial changes in the glottal pulse spacings, as observed in naturally-occurring irregular phonation. In our experiences, this can be achieved by removing one or two consecutive cycles (and thus doubling or tripling that specific fundamental period), without the need for fine control over pulse positions.

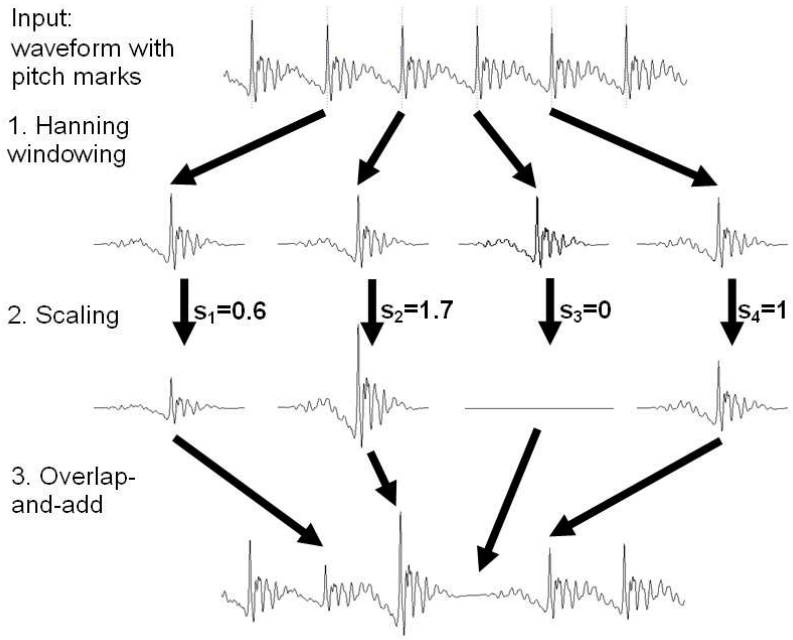


Figure 2. Illustration of the transformation method that introduces irregular pitch periods into the speech signal.

Attenuating or zeroing an impulse response also scales down the background noise present during that fundamental period. For example, if several consecutive cycles are removed from a recording with audible background noise, the lack of noise in the transformed region might decrease the perceived naturalness. In order to avoid this problem, background noise (windowed out from the end of the recording, for instance) can be added to attenuated and zeroed impulse responses.

To transform a modal recording so it is perceived as rough, one should create a pulse pattern (glottal pulse spacing and amplitudes) characteristic of natural irregular phonation. To reach this goal, the scaling factors can be modeled after a sample region of natural speech with irregular pitch periods. The factors need to approximately match the irregular pulse pattern in that sample. These factors then can either be set by hand using trial-and-error or 'copied' as a pattern from the model recording.

When the scaling factors are set by pattern copying, one has to select both the regular region to be manipulated in the signal and the irregular region to be copied.

Then a 'stylized' pulse pattern is extracted from the irregular region, consisting of the s_i scaling factors to be used in transforming the regular region (i.e., not the absolute pulse positions and amplitudes). Note that the concept of stylized pulse patterns is somewhat similar to Malyska's (2008, p. 29-40) glottal event patterns.

The stylized pulse pattern is initially constructed as a vector containing the relative amplitudes of the glottal pulses in the irregular sample. The amplitude of each period is measured as the peak amplitude (either positive or negative) around the pitch mark. The values in the stylized pulse pattern are expressed relative to the mean amplitude of L_A regular periods preceding the irregular region. When an irregular cycle is substantially longer than a reference cycle length (e.g., two or three times or more than the reference $T0_{ref}$, that is calculated as the mean of L_{T0} preceding regular cycles) zeros are inserted into the stylized pulse pattern since periods need to be removed from the regular recording at these points. The number of zeros to be inserted between two consecutive scaling values (i.e., the number of periods to be removed) is determined by the rounded ratio of the actual cycle length to the reference cycle length. Cycle lengths are measured as time differences between consecutive pitch marks. The number of periods used to calculate the reference cycle length (L_{T0}) and reference amplitude (L_A) is 5 by default, but can be set as a parameter. The scaling factors comprising the stylized pulse pattern are finally applied to the selected region of the waveform to be transformed.

A detailed description of the transformation method can be found in Bőhm et al. 2008 and in Bőhm 2009. Results of the evaluations presented in these publications illustrate that this transformation method reproduces (to some degree) most of the well-known acoustic characteristics of irregular phonation, and that listeners perceive the output to be acceptable as rough, natural-sounding speech.

3 About the program

A graphic tool, named *Glottalizer* has been developed to allow fast and convenient application of this transformation method. It runs in Windows and it is freely available for non-commercial use. The graphical user interface was programmed by the second author, while the transformation functions were implemented by the first author.

The program provides a means for a) the parallel display of both the waveform to be modified and the model waveform; b) copying stylized pulse patterns; and c) convenient iterative refinement of the scaling factors, because the effects of the parameter changes are immediately visible and audible. The program also has the usual sound display and play functionalities, as well as a command history.

Figure 3 shows a screenshot of the program in operation. The bottom panel displays the waveform of the recording to be manipulated. The top panel depicts the model waveform that can be used to guide the irregular phonation transformation (either manually or by copying its pulse pattern); a model recording that contains irregular phonation can be loaded into this panel. Note that the model recording cannot be manipulated. In order to open a wave file in either one of these two panels, a corresponding pitch mark file must also be available (such files can be

ganarated by, for example, Praat, in PointProcess format). The pitch marks can be overlaid on the waveform and can be edited and saved.

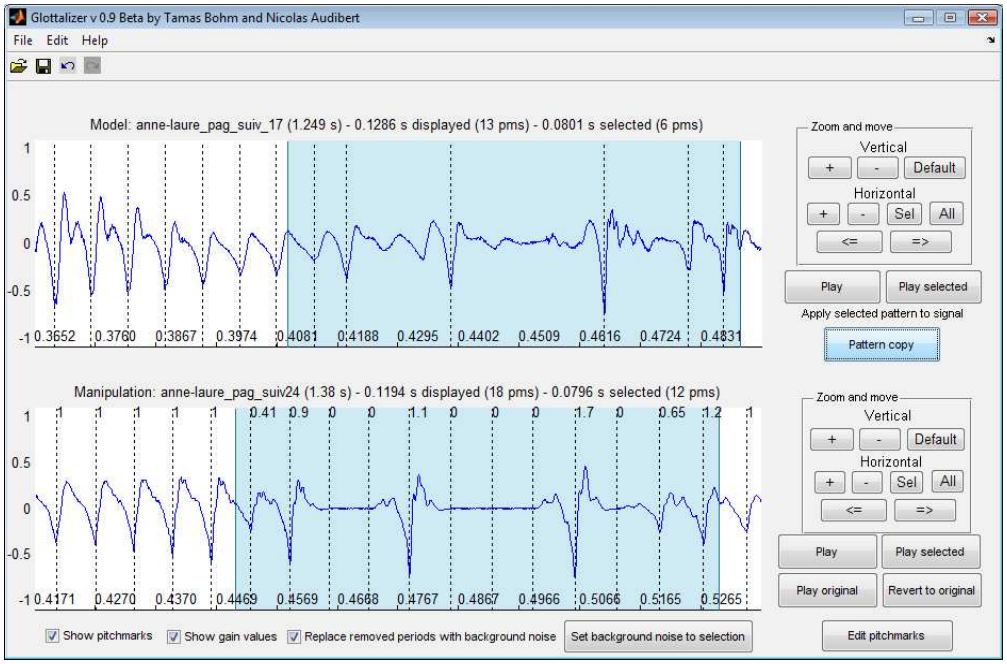


Figure 3. The screen of the program after copying the pulse pattern in the selected region of the model recording (top panel) to the other waveform (bottom panel); scaling factors are shown above the pitch marks (dashed lines).

In the bottom panel, individual periods can be scaled, removed (scaled to zero) or reverted to their original form (i.e., resetting the scaling factor to 1) by a simple mouse click. The applied scaling factors are shown above the manipulated waveform and can be saved to a separate file that can be reloaded later.

The transformation can also be carried out by copying a stylized pulse pattern. In order to do this, one has to select the region in the model waveform that has the target pulse pattern, and the region in the bottom panel where the pattern is to be applied. To enable pattern copy, there should also be enough pitch marks preceding the model selection to calculate the reference values.

A more comprehensive documentation is available on the Glottalizer website.

4 Acknowledgements

The authors are grateful for Véronique Aubergé, Stefanie Shattuck-Hufnagel and Géza Németh for their advice and support. The first author was funded by the ETOCOM project (grant TAMOP-4.2.2./08/1/KMR from the National Office for Research and Technology) and the NAP project (grant OMFB-00736/2005 from the National Development Agency).

References

Bőhm, T. 2009 *Analysis and modeling of speech produced with irregular phonation.* Ph.D. dissertation, Budapest: Budapest University of Technology and Economics.

Bőhm, T., Audibert, N., Shattuck-Hufnagel, S., Németh, G. & Aubergé, V. 2008 Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles. In *Proceedings of Acoustics'08.* 6141-6146.

Dilley, L., Shattuck-Hufnagel, S. & Ostendorf, M. 1996. Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics,* 24(4): 423-444.

Gobl, C. & Ní Chasaide, A. 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication,* 40(1-2): 189-212.

Henton, C. & Bladon, A. 1987. Creak as a sociophonetic marker. In Hyman, L. M. & Li, C. N. (eds.): *Language, speech and mind.* London: Routledge. 3-29.

Laver, J. 1994. *Principles of phonetics*. Cambridge: Cambridge University Press.

Malyska, N. 2008. *Analysis of nonmodal glottal event patterns with application to automatic speaker recognition.* Ph.D. dissertation. Cambridge: MIT.

Moulines, E. & Charpentier, F. 1990. Pitch-synchronous wave-form processing techniques for text-to-speech synthesis using diphones. *Speech Communication,* 9(5-6): 453-467.

Surana, K. & Slifka, J. 2006. Acoustic cues for the classification of regular and irregular phonation. In *Proceedings of Interspeech 2006.* 693-696.

Tamás Bőhm
Department of Telecommunications and Media Informatics of
Budapest University of Technology and Economics,
Institute for Psychology of Hungarian Academy of Sciences
Budapest, Hungary
e-mail: bohm@tmit.bme.hu
and
Nicolas Audibert
GIPSA-lab, Speech & Cognition Department (ICP), UMR 5216
CNRS/INPG/UJF/Université Stendhal, Grenoble,
Laboratoire de Phonétique et Phonologie, UMR 7018
CNRS/Université Paris 3,
France
e-mail: nicolas.audibert@gmail.com

On line acoustic features of segmental level speech data (Hungarian)

http://fonetika.nytud.hu/cvvc

This is a spoken word database for the acoustic presentation of CV, VC, VV, VVV, CC, CCC and CCCC sound combinations (2007)

1. search;
2. put the word into the basket for acoustic details.

Research and design

Gábor Olaszy, Research Institute for Linguistics, Hungarian Academy of Sciences (2005–2007).

Development

Database construction and programming: Kálmán Abari, University of Debrecen, Hungary. Collection of words and preparation for demonstration: Csaba Zainkó, Géza Kiss, Gábor Olaszy, Budapest University of Technology and Economics, Hungary.

Reference

Olaszy Gábor: Mássalhangzó-kapcsolódások a magyar beszédben (= Consonant clusters in Hungarian speech). Tinta Kiadó, Budapest, 2007.

Support

This research and development was supported by Hungarian Research Fund OTKA T0498456.

Remarks

Remarks can be sent to olaszy.gabor@gmail.com and abarik@delfin.klte.hu.

Presentation of Hungarian sound combination in speech

The aim of this speech database is to show the acoustic structures of Hungarian speech sounds and their combinations (and indirectly to show the acoustic results of the coarticulation).

The structure of the database

The sound combinations are shown in spoken words. Every possible form is presented for CV,VC,VV and CC combinations. The others, as VVV, VVVV and CCC, CCCC are represented by the most frequent items. The words are produced by a male and a female speaker while reading a word list.

Sound symbols

The special sound symbols used in the phonemic descriptions can be seen in Figure 1. The j + sound symbol means a hiatus resolution in certain VV combinations. The symbols and their counterparts in IPA are shown in the “sound table” as well.

A: = [a:], a = [ɔ], o = [o], u = [u], ü = [y], i = [i], E: = [e:], O = [ø], e = [ɛ]
b = [b], d = [d], g = [g], G = [ɟ]
p = [p], t = [t], k = [k], T = [c]
m = [m], n = [n], N = [ɲ]
j = [j], h = [h]
v = [v], z = [z], Z = [ʒ]
f = [f], s = [s], S = [ʃ],
dz = [dz], dZ = [dʒ]
c = [ts], C = [tʃ]

Figure 1. The special sound symbols used in the phonemic descriptions

How to search in the database

Two ways are available for searching sound combinations.

a) Speech sound search (Figure 2: upper left box). Here, the sound combination can be given directly (please use the “sound table”, Figure 3). IPA symbols show the sound.

b) Letter-based search (Figure 2: lower left box). Here, the letter sequence can be given and letters will be searched. Attention! Hungarian letters must be used.

The result of the search

A list is given below (Figure 4), containing the words found (every word in a separate box). Every word box contains the textual form, the sound representations of the word, the sound durations and three command words in brackets above the sound symbols. The three command words in brackets are: details in sound wave [Det.], baskets for acoustic presentations are in [Bas.1] and [Bas.2].

Acoustic presentation

Two forms of acoustic presentation are available.

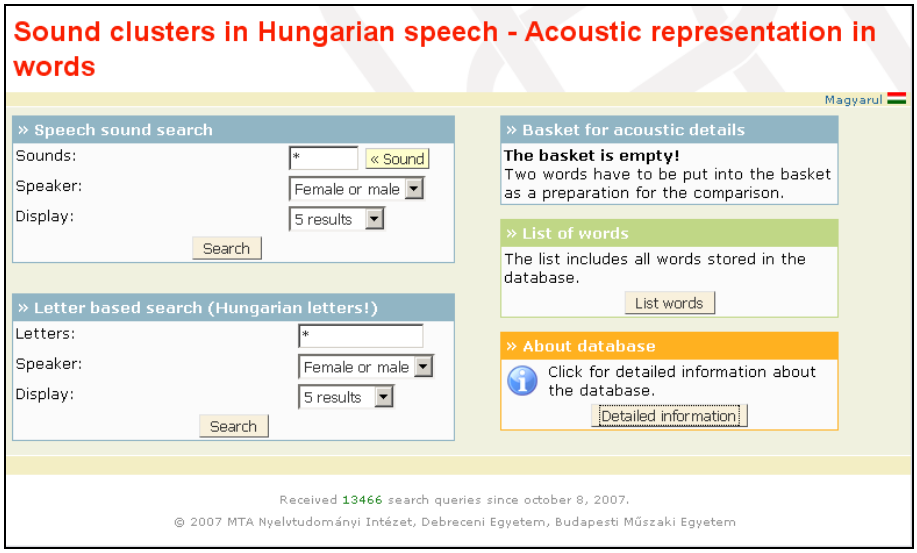


Figure 2. Speech sound search

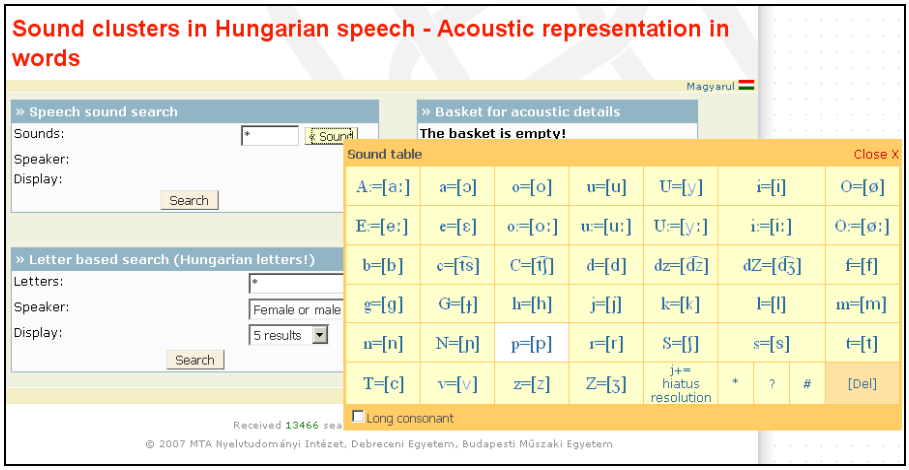


Figure 3. The “sound table”

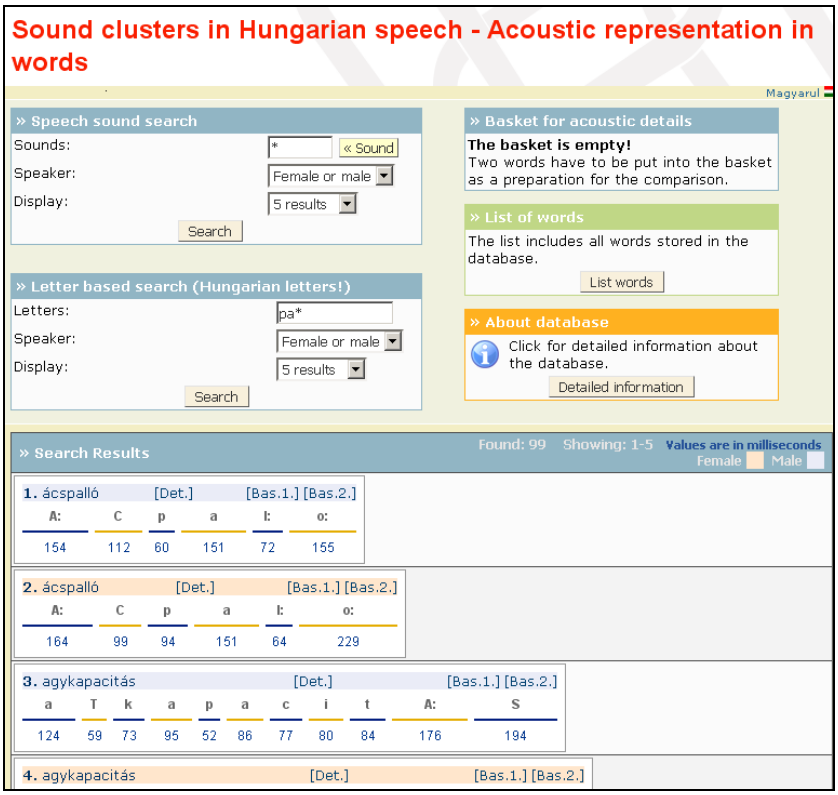


Figure 4. The result of the search

Wave form in detail [Det.]

Click on [Det.]. The waveform of the word will be shown as a separate window (Figure 5). Sound boundaries can be shown inside the waveform as well. Zoom facilities are available with parallel audio demonstration. Sound duration measurements (fine scale) can be done in this picture as well. The wave form can be played with normal, slower, or faster speeds. In addition, repeated play can be asked for (e.g. for one period of a voiced sound).

Acoustic structures and comparisons - [Bas.1], [Bas.2]

You have to put the word you want from the search list into the basket (for example click on [Bas.1] and next on [Bas.2], as a result, the selected word will be placed into the basket). Two items always have to be put into the basket! Press the "Acoustic presentation" and the result will be shown in a separate window (Figure 6). The following acoustic presentations are available: waveform with sound boundaries, spectrogram, or intensity. Listening to the whole word or part of it is also available. On a selected part of the waveform, duration measurements can be performed and the results are shown in ms. The same can be done on spectrograms (frequency is measured and displayed in Hz in the small window) and on intensity curves (the dB value is displayed in the small window).

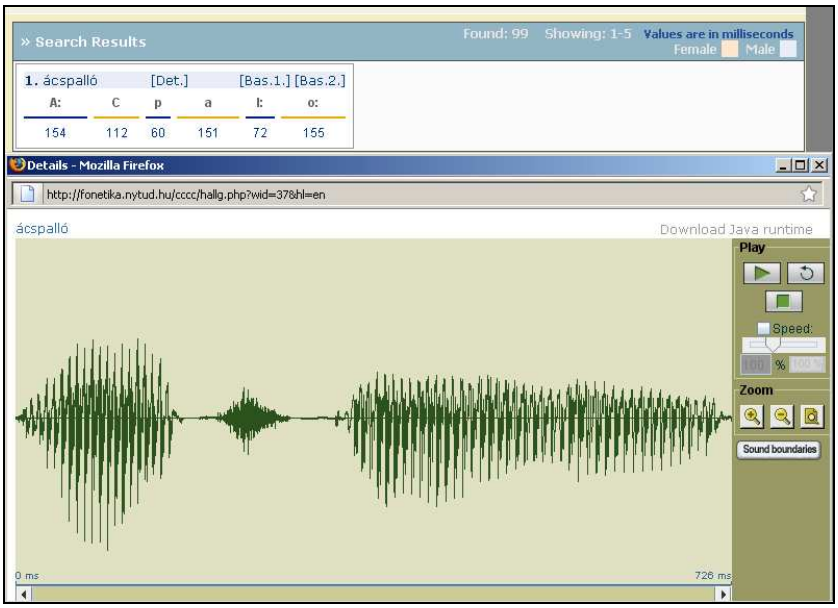


Figure 5. The waveform of the word

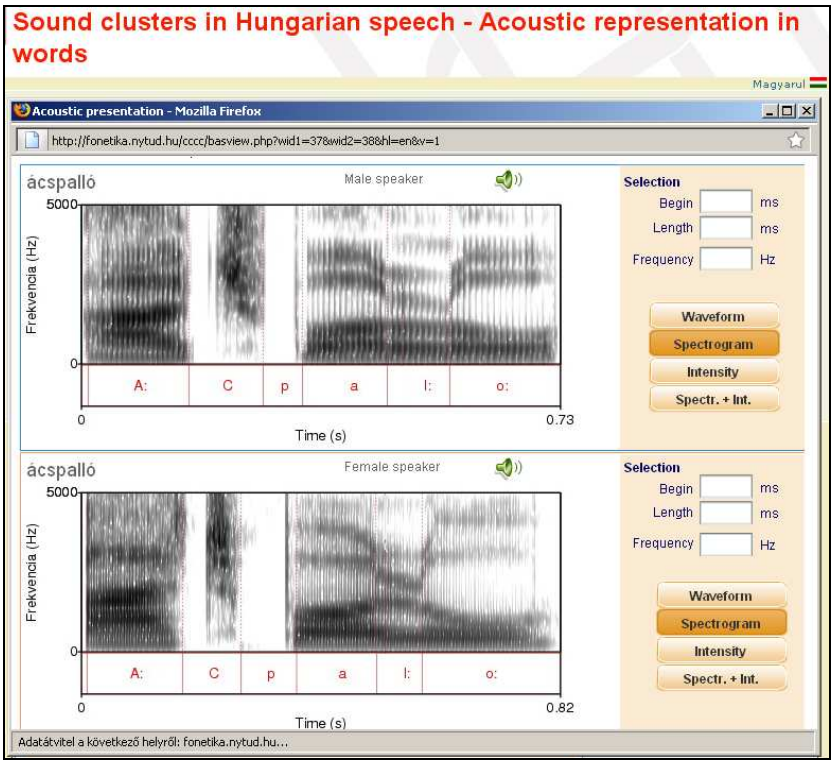


Figure 6. Acoustic representations

List of words

The list of words are shown used in the database. Clicking on a word results in a search command according to letters.

Gábor Olaszy
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics, Hungary
e-mail: olaszy@tmit.bme.hu

From text-to-sound - A Hungarian word level pronunciation database using IPA symbols

1 Introduction

Humans perform letter-to-sound conversions automatically during the course of reading. The rules of the mother tongue are applied for most words, however, the pronunciations of loanwords, names (surnames, geographical names, brand names) and other foreign words and expressions may cause problems because the spelling of the language of origin is retained. To help people in these instances, classical, printed pronunciation dictionaries are used.

Electronic pronunciation dictionaries may significantly differ from their classical printed counterparts. The methods used to compile the lexicon, the size of the compilation, as well as the functions of the dictionary are all determined by the range of possibilities provided by information technology. Electronic pronunciation dictionaries are essential for speech technology applications and also for linguistic research.

To date, no electronic pronunciation dictionary has been constructed for Hungarian to provide general information about the language. Our project is designed to rectify this situation. We will briefly survey our work related to the electronic collection of lexical items, the implementation of the data set, and outline the possibilities for its application.

The main feature of Hungarian is its agglutinative nature, making Hungarian morphology rather complicated. The estimated number of Hungarian word forms is close to 1 billion. The Hungarian writing system uses an expanded Latin alphabet. This orthography of this language is largely transparent, i.e. the pronunciation of the majority of words can be derived from the orthographic form.

2 Material and method

The structure of electronic pronunciation dictionaries is basically simple: they provide the orthographic form, and in the same row, the phonetic transcription of the given sequence is provided with sound symbols. In addition, there are several possibilities for accessing this large and well organized database (i.e., grouping,

running statistics etc.). The rendering of pronunciations is not a trivial task, even in Hungarian. One can make an effort to provide a deep, scientifically based transcription or an easily readable one. The former one must contain the sound symbols for all possible speech sounds of the language (i.e., the phoneme representations and the allophones); the latter can use other sound representations. Our goal in developing this system was to construct a pronunciation database which represented the pronunciations possible in the language (and to avoid constructing exceptions to account for foreign vocabulary items). This system uses IPA symbols for all Hungarian sounds. In addition, the easier readable form is also represented as a choice, i.e., the person can ask for the use of Hungarian letters in the transcription (where the allophones are not distinguished from their original phoneme representative). Text material was collected from the Web, lexical items (different forms) were derived from the texts, the phonetic transcription were performed using software (based on rules) and manual manipulations. Testing of the database occurred over a one year period. The accuracy of the text-to-sound conversion was higher than 99%.

3 The development

The development of this electronic pronunciation dictionary proceed through four major steps: (1) the selection of texts from the Web, (2) the compilation of the items for the rough database, (3) the screening of lexical forms and the compilation of the final word form database, (4) the definition of the pronunciation for each item, and (5) the testing for transcription accuracy of the entire database. The development of this dictionary took 4 years.

3.1 The selection of texts from the Web

The text material was compiled by means of automatic methods, relying on a large, electronically recorded text corpus collected automatically from the Internet (Zainkó & Németh 2001). Online editions of newspapers and the collections of e-libraries also were downloaded. We chose to use online editions of newspapers and e-libraries to collect data because, in our experience, they contain more carefully edited and revised texts than ordinary websites. Despite such a selective way of choosing our sources, we had to screen out certain material, like that written in a foreign language on these sites. The detection of non-Hungarian texts was carried out by means of a language detection software we developed. Language detection took place predominantly at the sentence level. Sentences written in languages other than Hungarian were eliminated from text storage. If the number of sentences written in a foreign language outnumbered the sentences written in Hungarian, we left the text out of the corpus. The final text corpus consisted of 80 million words and formed the basic text material for the further work.

3.2 The compilation of the lexical items for the database and screening of word items

In the 80 million-word text corpus, we found approximately 1.5 million different word forms, which then constituted the searchable textual base of the pronunciation

dictionary. The definition of word form was as follows: **a lexical unit in a text, which is delimited by non-letter characters** (mainly spaces). The letter-content in each and every word form differs from every other word form of the dictionary by at least one letter. A second screening was performed on the selected word forms to delete lexical units which do not belong to Hungarian. At the end of this screening, a searchable Hungarian lexical unit database (word forms) was formed, which served as the orthographic part of the future pronunciation dictionary. Further screening was conducted continuously until the end of project development. If a non-Hungarian element was found, it was deleted.

3.2.1 Statistical representation of the database

Word forms occur at a given rate of frequency in the overall text corpus. If we take this particular rate of frequency into consideration, we can make a coverage diagram that shows what percentage of the whole text corpus is covered by word forms (see Figure 1). This diagram indicates that the most frequent 1.5 million word forms cover 99.9% of the 80 million words corpus.

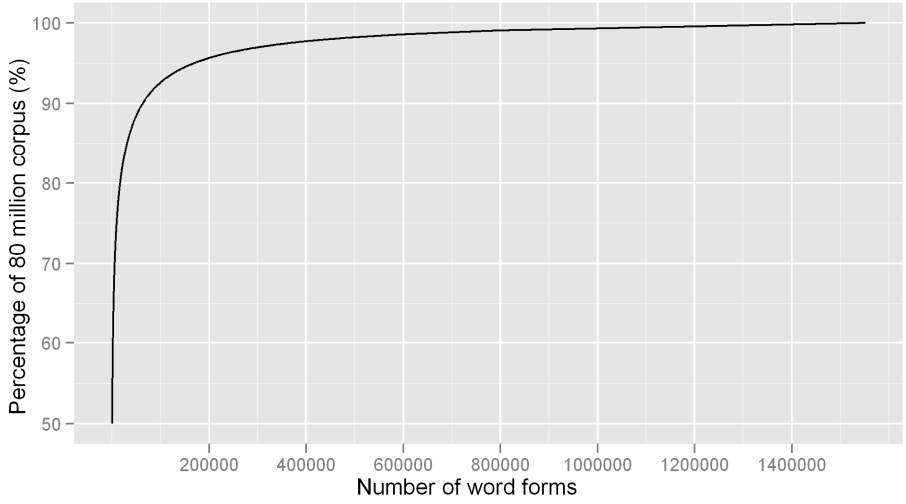


Figure 1. Cumulative percentage of the 80 million text corpus that is covered by the most frequent word forms. Word forms are in order of decreasing frequency along the x-axis.

We have carried out a statistical survey of the word forms in the dictionary. These results will be displayed in terms of syllable and stem distributions. We examined the most common word forms based on how many syllables used. This distribution is shown in Figure 2. Our survey revealed that the most common word forms in Hungarian consisted of 3, 4, 5 and 6 syllables (note that this distribution pertains to word forms, and not to the general distribution of words occurring in Hungarian texts). Detailed information about the number of syllables in Hungarian words can be found in (Szende 1973). These word stems were examined using the hunmorph

toolkit. Details on this open source morphological analyzer can be found in Trón et al. (2005, 2006). Up to 91.500 stems (6.21%) have been counted in the 1.5 million database.

3.3 Transcription of the lexical items into sound symbols

A transcription procedure was applied to the collected group of word forms. This consisted of a basic letter-to-sound conversion. At this point, two important issues had to be clarified: the selection of the applied phone set and the method of transcription.

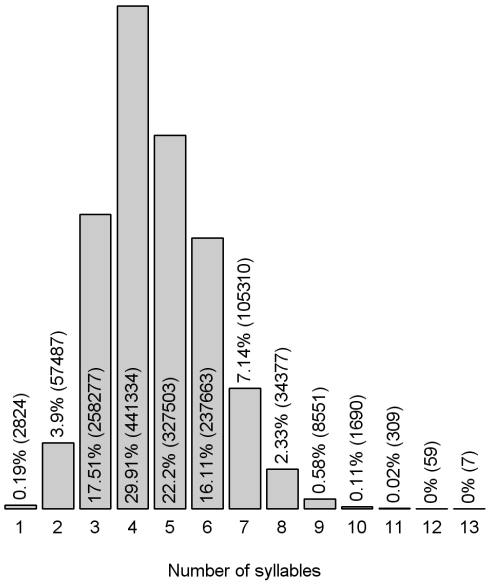


Figure 2. Distribution of word forms in the pronunciation dictionary by number of syllables

3.3.1 Selecting the phone set

The first important question addresses the kind of phoneme or phone set that should used to encode pronunciation. This, on the one hand, means determining the level (or depth) of the analysis and, on the other hand, the choice of the phonemic or phonetic alphabet to be used. Following our aim a precise transcription was performed and the IPA symbols were used for the phonemes and for the allophones noted (Taylor 2009). The phone set of the transcription consists of 14 vowels (short and long sounds), 25 short consonants and 25 long ones, 8 symbols for the allophones and one symbol for marking of any long sound (Table 1).

3.3.2 Transcription forms

In the database, one can choose from three formats to represent the pronunciation of the given lexical item:

• International phonetic symbols (IPA) (e.g., *küldte* = [kyltɛ] ‘sent’, *megkap* = [mɛkːɔp] ‘get’). IPA symbols provide unambiguous pronunciation for native speakers of any language, thus they are language-independent.

• Traditional Hungarian transcription (the pronunciation is given by Hungarian letters, e.g., *küldte* = “külte” ‘sent’, *megkap* = “mekkap” ‘get’).

• ASCII-based sound code symbols (TMIT) developed in Hungary, mainly for internal sound representation in computer programs (e.g., *küldte* = “kUlte” ‘sent’, *megkap* = “mek:ap” ‘get’). It is used mainly in applications of speech technology.

Table 1. The transcription between the orthographic characters and phonemes in IPA and TMIT symbols. Cells in the columns: Orthographic symbol | IPA-symbol | TMIT-symbol | Hungarian example.

Vowels

a	ɔ	a	*ad*
á	aː	A:	*áll*
e	ɛ	e	*el*
é	eː	E:	*él*
i	i	i	*irt*
í	iː	i:	*ír*
o	o	o	*ok*
ó	oː	o:	*ól*
ö	ø	O	*öl*
ő	øː	O:	*őt*
u	u	u	*un*
ú	uː	u:	*út*
ü	y	U	*üt*
ű	yː	U:	*űl*

Consonants

b	b	b	*bab*
p	p	p	*pap*
d	d	d	*dal*
t	t	t	*tél*
gy	ɟ	G	*gyár*
ty	c	T	*tyúk*
g	g	g	*gép*
k	k	k	*kép*
m	m	m	*mér*
n	n	n	*nép*
ny	ɲ	N	*nyár*
j, ly	j	j	*jön, lyuk*
h	h	h	*hír*
v	v	v	*vár*
f	f	f	*fal*
z	z	z	*zaj*
sz	s	s	*szél*
dz	d͡z	dz	*bodza*

Consonants cont.

c	t͡s	c	*cél*
zs	ʒ	Z	*zseb*
s	ʃ	S	*seb*
dzs	d͡ʒ	dZ	*dzsem*
cs	t͡ʃ	C	*csel*
l	l	l	*lép*
r	r	r	*rét*

Allophones

-	ʲ	j+	*fia*
-	ç	J	*lépj*
-	x	H	*doh*
-	x́	CH	*pech*
-	ŋ	n+	*ing*
-	ⁿ	n'	*unsz*
-	a	A	*sztrájk*
-	ɛː	e:	*khmer*

3.3.3 Grapheme-to-sound symbol conversion

To generate a correct canonical string of sound symbols from the written form, one can choose from a variety of methods (Schultz – Kirchhoff 2006), ranging from manual transcription (using a -crafted dictionary) to automatic coding (using an automatically generated dictionary). For relatively small vocabularies (e.g., exception dictionaries), the first approach is viable, however, for larger dictionaries, the second method is the most feasible. In the present case, a combination of the above approaches was used to increase the pronunciation accuracy of word forms. An overview of the variety of published grapheme-to-phoneme conversion techniques can be found in Taylor (2009), and in Bissani & Ney (2008). For the

implementation of the introduced Hungarian grapheme-to-phone (G2P) database, an algorithm had to be devised whose chief elements involved the rules of pronunciation specified in Hungarian (Olaszy et al. 2000). The basic idea of the algorithm was: there are rules and every rule has a list of exceptions. The question in each case was what to designate as a rule. The following principle was applied: the prevailing number of occurrences represented the rule, while the smaller number of occurrences was regarded as the exceptions.

3.3.3.1 Grapheme-to-phone rules

In the following list, examples are given for the rules and their exceptions.

• The same grapheme can represent different phones in different contexts (e.g., /t/ can represent [t] in the word ***t**alál* 'he/she finds', or [d] in the word *kú**t**ban* 'in the well'.

• The same phone can correspond to different graphemes in different contexts (e.g., [t] can be represented by /t/ in the word ***t**alál*, or /d/ in the word *pa**d**ka* 'berm').

• A sequence of two graphemes can represent a single phone. For example *te**lj**es* [tɛ**jː**ɛʃ] 'complete'.

• The same sequence of two graphemes, can represent different phones depending on the syllable and morpheme boundaries of the lexical item. For example in word *te**lj**es* [tɛ**jː**ɛʃ] and *fe**lj**avít* [fɛ**l**jɔviːt] 'upgrade'.

• Some graphemes do not represent a phone. A mute [h] is noted in the word *méh* [meː] 'bee', or *juh* [ju] 'sheep', in absolute final position (but in case of having a suffix, the [h] appears, like *juhok* [juhok] 'sheep plural'.

• The hiatus resolution cannot be represented by any grapheme, like *fiú* [fiʲuː] 'boy'.

• Shortening of sound. Some grapheme sequences imply shortening in pronunciation, like *jobbra* [jo**b**rɔ] 'to the right'.

• Lengthening of sound. Some letters representing a short sound will be pronounced as a long sound, like *lesz* [lɛsː] 'will be'.

3.3.3.2 Exceptions in transcription

To transcript all items of the presented text-to-phone database accurately, an exceptions list was used. This procedure accommodated foreign, borrowed, or other types of words, abbreviations, and every other form of pronunciation which could not be covered by the first two modules (*city* = [siti], *plaza* = [plaːzɔ], *Peugeot* = [pøʒoː]. In the dictionary, the determination of the pronunciation of Hungarian family names also can be found (*Kossuth* = [koʃuːt], *Bernáthffy* = [bɛrnaːtfi], *Előrsy* = [øːrʃi]), *Unghváry* = [ungvaːri]). The exceptions dictionary contains a total of 12.000 items.

3.4 Storage of the database

Pronunciation dictionaries can be stored in a simple dictionary format, which is easy to read and process both for humans and computers. Every line of the dictionary format begins with the orthographic form followed by the pronunciation. Dictionaries with large lexicons store these data in relational databases for quick search. This method of storage enables the recording of other types of information,

such as word class, waveforms for listening, etc. The latter form was applied to the Hungarian database as well. The relational database consists of 1.5 million entries and occupies approximately 150MB of storage space. In addition to the orthographic and pronunciation forms, two smaller word type groups were defined (family names and names of all the towns and villages in Hungary), in addition to links to 60000 audio WAV sound files.

3.4.1 Audio service

Speech technology solutions enable the incorporation of a large number of audio examples into a pronunciation dictionary. This helps the user to listen to the real pronunciation, to feel the sound durations and the rhythm of the lexical item (in Hungarian pronunciation, keeping the short – long opposition is very important). By listening to the items, international users can make a parallel comparison between the sound symbols and the spoken sound. There are about 60,000 items in the introduced Hungarian pronunciation.

3.5 Verification

An important aspect of any large database is its accuracy. If the pronunciation of a word is stored incorrectly in the dictionary, efficiency is affected. To verify the 1.5 million items in the database, the following procedure was performed. Lists were generated by linguists. These lists were examined and incorrect items were corrected. The process was repeated multiple times. As a result, we identified several error categories:

• spelling errors in word forms,

• pronunciation of foreign words, acronyms, and word forms with unexpected pronunciations which were not accounted for in our exception dictionary, e.g., *Clemenceau, Glasgow, sombrero, UNESCO,*

• ambiguity in composed word boundaries, like *pénzeszsák* without morphological segmentation can be *pénzesz|sák** or *pénzes|zsák.*

• distinctive cases in forming or not forming hiatus resolution, like *kiabál* [kiʲɔbaːl]‘shouts’ opposition to *kiállítás* [kiaːlːiːtaːʃ] ‘exhibition’.

Pronunciation dictionaries require maintenance, so its development cannot be conclusively terminated, though maintenance costs and human resources may be reduced. Language changes, newly coined words emerge, others get transformed, and in the case of certain new linguistic items, it is necessary to determine the forms of pronunciation. The continuous maintenance of this Hungarian pronunciation database will be organized in the near future.

4 Presentation of the database

The presented Hungarian pronunciation dictionary contains three marked groups of 1,475,391 word forms: general items (1,471,295), family names (743), names of Hungarian towns and villages (3,353). The total number of phones in transcripted forms is 16,482,254. The distribution of the number of phones is shown in Figure 3.

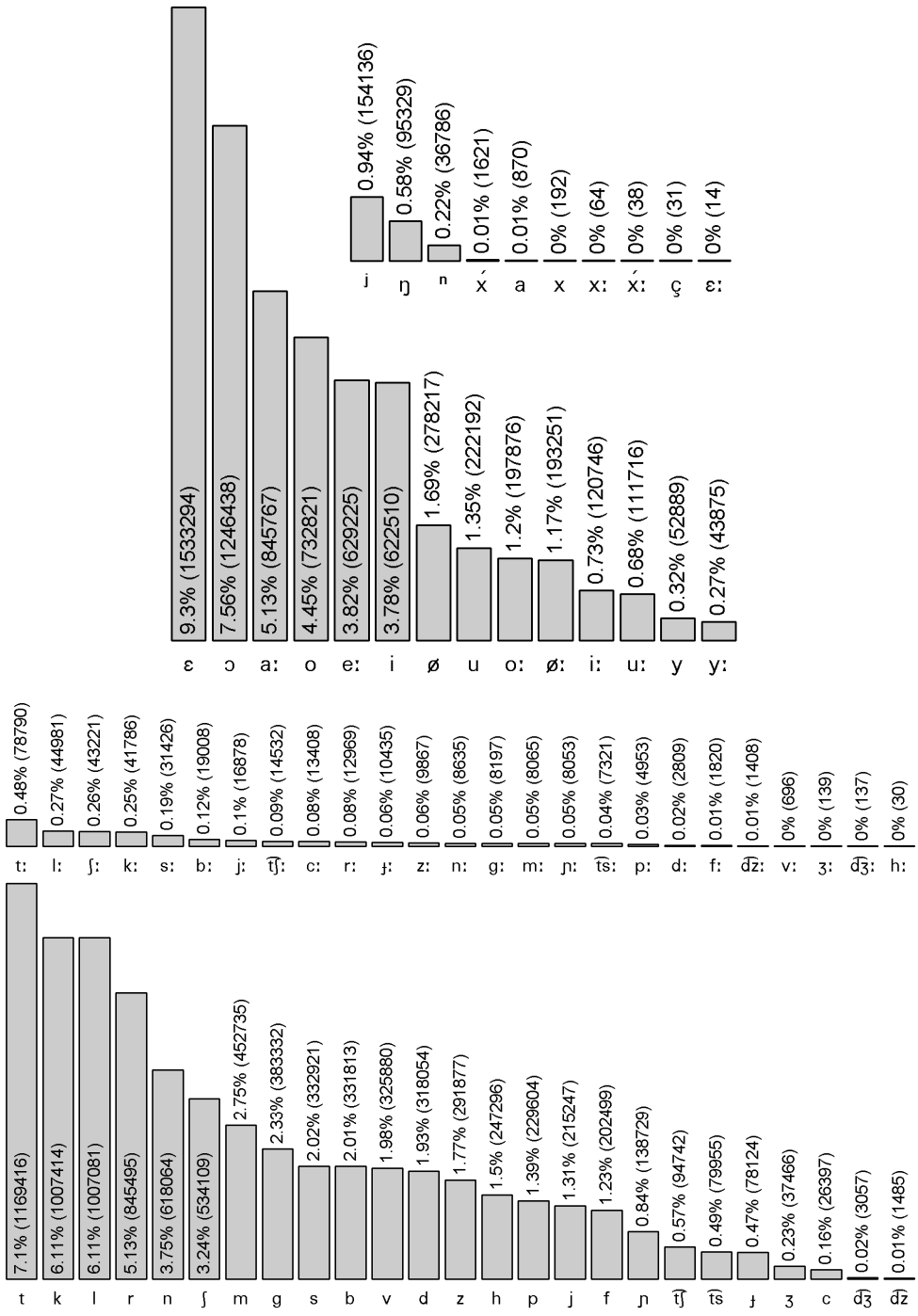


Figure 3. Distribution of phones in the database

The distribution of the number of sounds in one word form is shown in Figure 4. This curve shows that our database contains word forms typically consisting of 7-15 sounds (with most of them using 10 sounds).

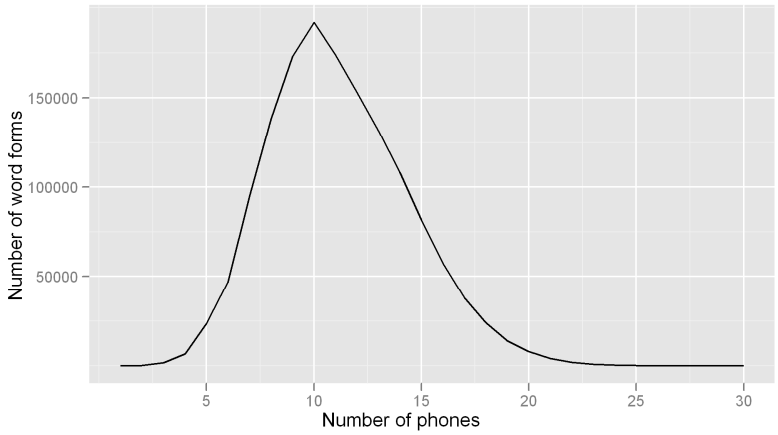


Figure 4. Distribution of the number of phones occurring in a word forms

The dictionary consists of a total of 15,006,863 diphones, made up of 2145 unique diphones. The corpus covers 78.57% of the theoretically possible diphones in Hungarian. Out of these unique diphones, 4.32% occur only once in the corpus. The coverage diagram shows what percentage of the diphone instances are covered by the unique diphones (see Figure 5). The most frequent 500 diphones cover 90.47% of 15 million diphone corpus. Table 2 shows the most frequent 20 diphones in order of decreasing frequency.

5 The use of the pronunciation database

Electronic pronunciation dictionaries can be used in many fields of research and development as follows:

Text-to-speech systems. Most text-to-speech systems use rule & exception approaches, a combination of G2P rules and an exception dictionary to perform pronunciations. Traditionally, the rules are the primary elements of the conversion and dictionary look-up is a secondary, which is only used in cases when the rules failed. However, this process is entirely reversible if a high-quality lexicon and enough storage space are available. Then, we can look-up items in a large dictionary first, and apply rules, when dictionary fails (Taylor 2009).

Automatic speech recognition systems. One of the main challenges in automatic speech recognition is pronunciation variation (Vicsi & Szaszák 2004). Pronunciation variation examinations require to produce canonical pronunciation of words and compare it to the label files obtained by audio-visual segmentation. The process may use look-up in a large dictionary as the primary strategy to determine the canonical pronunciation form of the utterance.

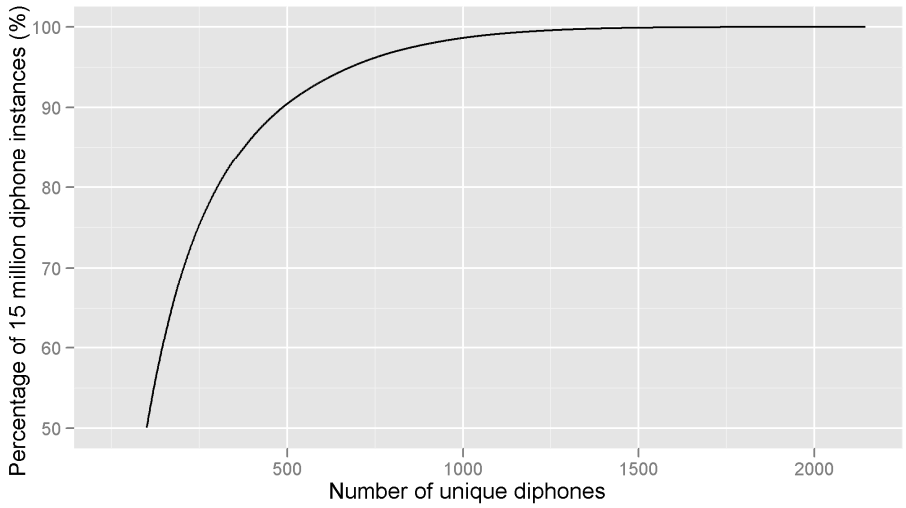


Figure 5. Cumulative percentage of the 15 million diphone instances covered by the most frequent unique diphone. Diphones are listed in order of decreasing frequency along the x-axis.

Table 2. The most frequent 20 diphones in order of decreasing frequency

Diphone	Number	Frequency (%)	Cumulative frequency (%)
ɛl	218186	1.45	1.45
ɔt	172414	1.15	2.60
ɛk	165752	1.10	3.71
ɛt	163182	1.09	4.79
tɛ	158274	1.05	5.85
ɔl	152060	1.01	6.86
ɛr	150991	1.01	7.87
aːʃ	143292	0.95	8.82
ɔk	142015	0.95	9.77
lɛ	130957	0.87	10.64
rɛ	118733	0.79	11.43
ɛn	114974	0.77	12.20
taː	114850	0.77	12.97
tɔ	112056	0.75	13.71
aːr	106032	0.71	14.42
mɛ	103170	0.69	15.11
rɔ	101808	0.68	15.78
or	98661	0.66	16.44
aːl	95857	0.64	17.08
eːʃ	92551	0.62	17.70

Grapheme-to-phoneme conversion. Similarly to the way G2P algorithms can be applied to the development of pronunciation dictionaries (see 2.3.2), pronunciation dictionaries can be utilized to develop G2P algorithms. All data-based approaches require a pronunciation dictionary that contains the orthographic as well as the phonemic and phonetic transcription of a large number of words. For example, we evaluated our dictionary using the method described in (Bisani & Ney 2008) and the publicly available Sequitur G2P grapheme-to-phoneme converter (Sequitur G2P 2008). Five different sizes of randomly selected word forms were used for training (1000, 5000, 10000, 50000, 100000), different n-gram models (1-6), and the same 100000 word forms were used for evaluation. Figure 6. shows the word error rates for different models (lines) by different training size (x-axis).

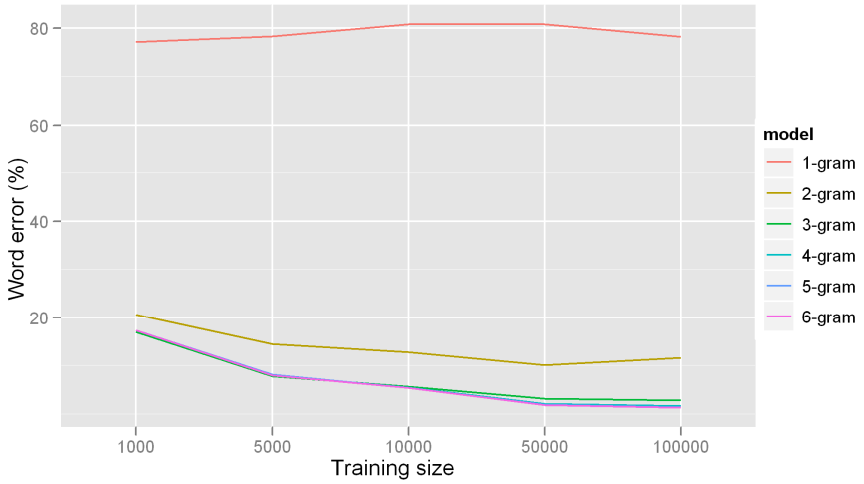


Figure 6. Word error rates referring to different n-gram models trained by different number of word forms

Education. From an educational point of view, a searchable database is useful for giving examples of any pronunciation rule and exception. Examples can be shown for the formation of different allophones, as well and the phenomena of hiatus resolution.

Research. This database can be used for any research and further development. As this is the first such database, it can serve as a learning platform for new, statistical text – sound conversion solutions. This database may be thought-provoking for linguists and mathematicians. The most attractive new result is the determination of phonetic level Hungarian. Analysis performed on the database revealed that Hungarian can be regarded as a phonetic language 74% of the time. This means that taking the whole set of letter sequences compared with the phone symbols, a difference between them occurs 26% of the time. The specificity of this information for Hungarian was not previously known.

Data mining. The database can help certain data mining research (i.e., how prefixing, suffixing works in Hungarian, letter sequences can be studied inside words, etc.).

New developments. Our dictionary provides pronunciations for many inflected forms and compound words. As a reasonably sized starting dictionary, it may be used to generalize conversions at morpheme boundaries to word boundaries to infer pronunciation of new entries of multi-word sequences.

References

Bisani, M. & Ney, H. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5): 434-451.

Olaszy, G., Németh, G., Olaszi, P., Kiss, G. & Gordos, G. 2000. PROFIVOX - A Hungarian professional TTS system for telecommunications applications. *International Journal of Speech Technology*, 3(3-4): 201-216.

Schultz, T. & Kirchhoff, K. (eds.) 2006. *Multilingual speech processing.* Orlando, USA: Academic Press.

Sequitur G2P 2008. *Sequitur G2P - A trainable Grapheme-to-Phoneme converter*. http://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html

Szende, T. 1973. *Spontán beszédanyag gyakorisági mutatói.* Nyelvtudományi Értekezések 81. Budapest: Akadémiai Kiadó.

Taylor, P. 2009. *Text-to-Speech synthesis.* New York: Cambridge University Press.

Trón, V., Gyepesi, Gy., Halácsy, P., Kornai, A., Németh, L. & Varga, D. 2005. Hunmorph: open source word analysis. In *Proceeding of the ACL 2005 Workshop on Software.* Ann Arbor, Michigan. 77-85.

Trón, V., Halácsy, P., Rebrus, P., Rung, A., Simon, E. & Vajda, P. 2006. Morphdb.hu: Hungarian lexical database and morphological grammar. In *Proceedings of LREC 2006.* Genoa, Italy. 1670-1673.

Vicsi, K. & Szaszák, Gy. 2004. Examination of pronunciation variation from hand-labelled corpora. In Sojka, P., Kopecek, I. & Pala, K. (eds.): *Text Speech and Dialogue: 7th International Conference (TSD 2004). Proceedings.* Springer. 473-480.

Zainkó, Cs. & Németh, G. 2001. Statistical text processing for automatic synthesis of speech. In *Proceedings of ECMCS2001 (EURASIP Conference on Digital Signal Processing for Multimedia Communications and Services).* Budapest. 644-647.

Kálmán Abari
Debrecen University, Hungary
e-mail: abari.kalman@gmail.com

APPLICATIONS USING SPEECH TECHNOLOGY RESEARCH RESULTS

Mindroom – Video news search system by voice

Due to technology advancements, more and more video and audio archives are available, not only for English, but also for less common languages like Hungarian, e.g., the audio archive of the Hungarian Parliament, the National Audiovisual Archive, archive of radio and TV stations, etc. The large size of these archives involves the problem of speech retrieval. Tagging the archive with some metadata is commonly used, but it is not sufficient for searching spoken stories. An archive is more valuable if it is searchable. The lifespan of the records is getting longer and additional services are being added to the database. Initially, these databases could be manually processed, but took at least 10 hours to transcribe one hour of speech. So, an automated solution is needed. In this article, we describe an LVCSR (Large-Vocabulary Continuous Speech Recognition) system for Hungarian, called VOXerver, and a video searching and tagging application called Mindroom, built with knowledge obtained from VOXerver. We provide a website, where anyone can search the most remarkable TV and radio news in Hungarian. Search indexes were created by speech recognition software.

1 Hungarian Large Vocabulary Continuous Speech Recognition (LVCSR)

A speech recognition engine which uses special acoustic and language is needed to recognize TV and radio news (a Hungarian LVCSR system). So, a recognition engine was developed at AITIA as a one-pass frame synchronous WFST decoder, called VOXerver. The acoustic and language models utilized are task specific and must be trained using domain-specific training data.

Acoustic modeling is largely independent from the recognition task, but acoustic environment and speech genre must be taken into account. Our goal was to recognize TV and radio news, which should be comprised of good quality, planned speech. More than 10 hours of news from different sources were collected and transcribed and added to our previously collected good quality speech database. These data were used to train acoustic models, which were standard HMM triphones.

In some cases, telephone interviews occurred during the news. Telephone speech differs significantly from studio-recorded speech, which typically results in significantly lower recognition accuracy. Therefore, a second acoustic model was trained to include the quality of telephone speech. Although we have trained both acoustic models, the automatic selection between them has not implemented in the current version of the news search system, but will be in the future.

For language modeling, we needed much more domain data than the 10 hours transcribed for the acoustic model training. Therefore more data had to be transcribed but a lot of news is published on radio and TV channel websites, which can be downloaded. Collecting a large corpus (~5.6 M words in our case) was the first step in training the language model. We needed some text processing, like converting numbers to their written form, eliminating punctuation, collecting acronyms, foreign words and proper names, converting non proper names to lowercase, etc. The cleared corpus contained more than 280,000 different word forms, and even with this large vocabulary the OOV (Out Of Vocabulary) rate was 3.6%. Most commonly used LVCSR systems apply words as basic lexical units. Word-based recognition of morphologically rich languages, however, can result in the above addressed problems: very large vocabularies, high OOV rate, and inaccurate language model parameter estimation due to the large number of distinct word forms. In the latter case, about 50% of the vocabulary occurs only once in the corpus.

These phenomena can be handled by changing the base units from words to sub-word lexical units, known as morphs. In this way, vocabulary size can be radically decreased and even OOV words can be recognized. Thus, recognition accuracies can be improved over the word baseline. Segmenting words into morphs is not trivial, however there are grammatical and statistical approaches to solve this problem. In our case, the statistical morph segmentation proved to be the most accurate. So, we segmented the training corpus into morphs, but the recognizer still has to emit word forms. Therefore, a special word boundary symbol is used in the language model, enabling the reconstruction of words from morphs.

Introducing morphs instead of words reduced the vocabulary by 90%, and reduced the word error rate by 10%. Now, about 80% recognition accuracy can be achieved in real time.

2 Mindroom – A searchable video archive

Mindroom is an application (http://www.mindroom.hu). The Hungarian news recognition engine was integrated into a practical tool, so audio and video files became retrievable by their real content, instead of by titles, tags and other manually added metadata, which are typically used.

If someone searches for a video related topic, like „International Monetary Found" (IMF), Mindroom can list all relevant videos (left side of Figure 1) from different sources. The recognition engine timestamps the resulting text, therefore one can watch any of the videos from the moment when the searched content was produced (right side of Figure 1), i.e., the user is not forced to watch the whole video, and search manually. Mindroom highlights the context of the searched expression in the videoplayer window, so seekers can easily find the relevant part of the video.

Visitors can only watch 30 seconds from a selected video, after that they are navigated to the original source of the content, i.e., the website of the TV station.

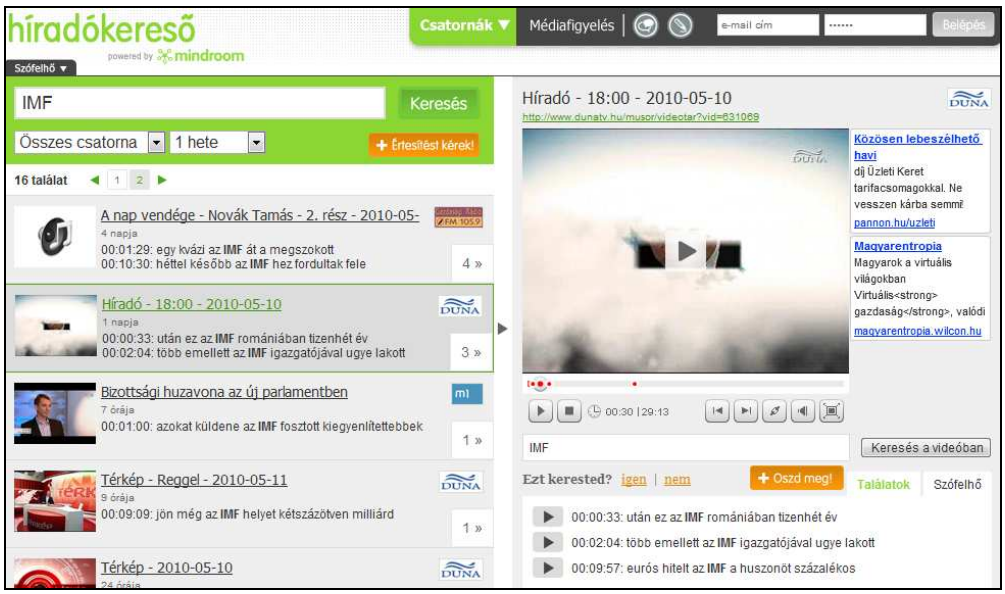


Figure 1. Mindroom – news searching webpage

3 Finding many Hungarian video news by only one search engine

The news indexing service was released in 2009. Since that time, Mindroom has been processing more than 13 hours of video and audio content every day, Mindroom is monitoring the most prominent Hungarian television and radio companies for economic and political news. More than 50 programs from seven TV and radio channels have been processed.

4 Automatic tagging – Fast overview

Mindroom's second most important function is an automatic tagging. It summarizes and visualizes the topic of the video in a tag-cloud (see Figure 2).

Figure 2. Tag-cloud of a video

Daily, weekly, monthly tag-clouds are summarized by Mindroom. It completes a statistical trend analysis to emphasize the most relevant keywords. In this way, Mindroom can provide a fast overview of the most important events in that time

period. Visitors can simply click a word, and Mindroom shows hit rates for stored videos.

5 Mindroom as a service

Large media factories have thousands of videos in their archives. Mindroom can make them searchable or can add metadata to this material using its API connections:

• Speech API: Media owner uploads their videos, and when Mindroom's process has been completed, the owner receives the time stamped recognition texts of the uploaded videos.

• Tagging API: Additionally Mindroom can create videos' tag-clouds and send them via Tagging API.

• Search API: If Mindroom receives a search expression from a third party website, it can give back relevant video ID-s, video titles, icons, time coded links etc. (This is useful, when someone wants to use Mindroom's search engine).

• Mindroom Embed: A Mindroom-Related Video engine can analyze article topics and match related videos from the Mindroom database to it. It will be displayed in an embedded window on a web page. As a result, news portal visitors can watch and play a short segment of videos related to the actual article, just like an embedded picture gallery (see Figure 3). The biggest advantage of this technology is that only an embedded code is needed to add a new and powerful feature to a website.

Figure 3. Mindroom related videos to articles by content

6 Business solutions

• Media monitoring: Mindroom can process all media related to business, economic, and political themes.

• Targeted ads in videos by content: Like Google adwords, Mindroom can analyze the content of ads and define relevant advertisements.

Tibor Fegyó
AITIA International Inc.
e-mail: tfegyo@aitia.ai
and
Sándor Somos
Digital Natives Ltd.
e-mail: somos.sandor@digitalnatives.hu

Spoken dialogue-based phone information system for pharmaceuticals

Automatic voice access of patient information leaflets

This is a report about an automatic information system, that uses speech synthesis and automatic speech recognition for a very limited area i.e. for a pharmacy. The name of the system is Medicine Line (MLN) and it has been operating in Hungary since December, 2006. The MLN system ensures 24 hour access to its information. The spoken dialogue input is processed by a specialized automatic speech recogniser (ASR) module (i.e., the caller says the name of the drug, the chapter title, etc. and the system recognises the voice). The output is given by a text-to-speech (TTS) synthesizer specialized to read drug names and medical texts, including Latin words correctly. The user can control the system by DTMF buttons too.

In Hungary, the National Institute of Pharmacy (NIP) coordinates the approval of new drugs and also their Patient Information Leaflets (PIL). Medicine Line reads the textual information of the PIL chapter by chapter to individuals. There are about 5000 different medications used in Hungary. New drugs enter the market regularly, and some are withdrawn after a certain time. The development of this system was supported by a GVOP National Project, and the EU as well.

1 Speech technology components in medical area

Using speech technology in medical areas is a developing industry. Radiologists can already dictate their analyses of X-rays, and their spoken words will be transcribed into text. Physicians can dictate prescription orders into wireless hand-held devices and the ASR enters the spoken items into text, which can be saved after confirmation into a central information system for later use (Vicsi et al. 2006). Investigations show that ASR – in spite of its present accuracy rate (80%-95%) – is being used more frequently in dictating medical reports (Grasso 2003).

One of the main problems of using speech technology in the medical field is having the software correctly process the special language elements (written or

spoken) of the medical field (Henton 2005), including its terminology (drug names, Latin words, etc.).

2 The MNL system

Two aims were taken into consideration when designing this pharmaceutical information system (http://www.gyogyszervonal.hu), as it is to be used mainly by the public.

The first was to allow 24 hour access to the patient information leaflet for every drug. The system can also help physicians and chemists in their work, by giving them quick access to the data (usually within 1 minute).

The second important aim was to design and develop a user friendly, easy to use system for all. That is why the spoken dialogue solution was chosen. A user independent specialized ASR module accepts the voice of the caller who can define the medication in question by giving its name. After recognition and confirmation of the drug name, the information leaflet of the drug is read by a TTS synthesizer which has been specially programmed to pronounce drug names, medical Latin words, and special expressions correctly. The dialogue between the caller and the machine is always provided by voice.

3 Drugs as a target

The National Institute of Pharmacy (NIP) coordinates the approval of new drugs in Hungary. The number of different medications used is about 5000. Pharmaceutical manufacturers submit a PIL about the drug, the text of which is approved by NIP. This leaflet contains information about the main use of the product, what is important to know before using this drug, how to use it, what side effects can occur, etc. New drugs come into practice regularly, and there are some which are withdrawn after a certain time, so the fluctuation of the products is rather high.

4 System components

The Medicine Line system has five main components; ASR for drug names, TTS for reading the text of PILs, dialogue controller, database and the automatic updater, which will be detailed below. (see Figure 1).

1. ASR for drug names

The main goals of the ASR for drug names are: to recognize the spoken drug over the telephone with more than 95% accuracy; to ensure an easy form to define new drug names; to recognize user commands with high accuracy while controlling the dialogue; and to provide flexible software support for the operator who supervises the system (maintenance, update). Updating is one of the most important activities of the system. Specialized grapheme to phonemes rules are applied in the automatic pronunciation model. So, the operator can manually control the phonemic transcripts of the drug names.

The acoustic models of the ASR module were taught on a SpeechDat-like Hungarian speech database, with approximately 20 hours speech (http://alpha.tmit.bme.hu/speech/hdbMTBA.php and http://alpha.tmit.bme.hu/speech/hdbtesztelen.php).

Three state left-to-right structure HMMs were applied with 10 mixture components per state. A speaker independent decision-tree was used to cluster cross-word triphone model states resulting approximately 2000 HMM states, which were trained using maximum likelihood estimations (Young et al. 2002). The front-end of the ASR is used to get MFCC-based acoustic features (Mihajlik et al. 2005).

Since users are forced to tell only the name of medication no word spotting or continuous speech recognition is applied. Therefore no language model, or in other words, a zero-gram model is used, resulting in a vocabulary of about 5K words. To speed up the recognition process, the recognition network was optimized off-line at the HMM state level using WFST algorithms (Mohri et al. 2002). One-pass decoding is performed using an enhanced version of the decoder described in another paper (Fegyó et al. 2003). The decoding is completed in close to real-time on a 3GHz CPU with a 32 channel load.

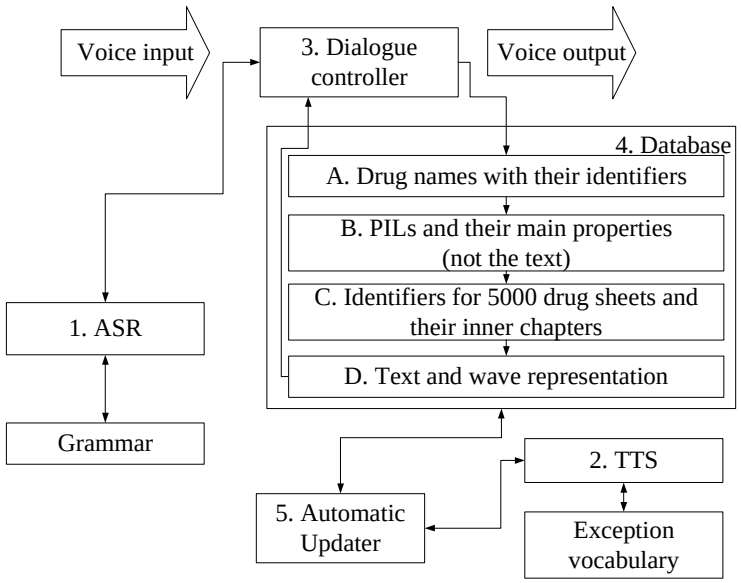


Figure 1. The main blocks of the Medicine Line system

2. TTS for pharmaceutical texts

The TTS module of the MLN system was developed from the Hungarian Profivox synthesizer (Olaszi et al. 2000). The Profivox-Med reading software has two special features beyond those utilized in Profivox. First, it is trained to pronounce thousands of drug names and Latin words correctly; ones that occur in the PILs. Second, it is prepared to synthesize (mainly from the prosodic point of view) the special style used by pharmaceutical manufacturers when writing PILs. It must be emphasized, that the texts of PILs are not composed for TTS conversion, but for human reading. This style is difficult and there are many spelling mistakes. Humans jump over the misspellings during reading and interpreting the text, but the TTS converts them

letter by letter, as they are written, so the sound will be incorrect. In our opinion manufacturers should be required in the future to take the speech technology demands in this field into account. This is their best interest.

Reading medical terminology by TTS. During the preparation of the Profivox-Med software for correct reading of medical terminology, all texts of PILs have been processed to pick out and collect the non-Hungarian words and text parts occurring in the PILs. The majority of the selected words were drug names and Latin words, but other items, like foreign words, abbreviations, chemical expressions have been found also (*N-hepa; hidroxi-propil-metil-cellulóz; 40 µg PGE; 800 mOsm/1; kallikrein inactivator unit; HMG-CoA; non-Hodgkin lymphoma*, etc.). This list has been examined and a rule based pronunciation sub-module was developed supported by a special exception vocabulary to reach the correct TTS conversion of these items. The rule-based part touched mainly the drug names and the Latin words (see Table 1). In this part, pronunciation rules have been defined for certain letter combinations, resulting in 243 rules.

Table 1. Example of pronunciation rules for drugs for TTS conversion

Drug name	Pronunciation	Word end rule
ACCOLATE	[ak:olat]	TE# = [t]
ACICLOSAN	[atsiklozan]	SAN# = [zan]
ACTILYSE	[aktiliz]	YSE# = [iz]
ALKA-SELTZER	[alkaseltser]	TZER# = [tser]

The exception vocabulary was comprised of transcription of the remaining words by sound symbols. The use of this combined solution in the TTS system resulted in the correct reading of medical terminology embedded in the text. The need to increase the vocabulary is continuous, because new drugs are constantly being developed and these new items cold be produced with the wrong pronunciation. In this respect, this part of the system is open and needs continuous manual work from the operator.

Special style of the PILs. Patient Information Leaflets are originally written for humans, not for automatic reading by a machine. As every profession, the pharmaceutical industry has a special style for these texts, developed independently of the manufacturer. In many cases, there are very long sentences, and their grammatical structure is complicated. A human reader understands the structure of the text, so the interpretation of the essential parts is not typically a problem. To read this by a machine, special pausing and an intonation strategy is needed in the TTS converter. Embedded medical expressions in the text also increase text difficulty.

In many cases, the sentences also have text parts between brackets (i.e., references to drug names, patient indications etc.). The application of special prosody is needed to make these text understandable by humans. Many sentences contain long lists. For example, side effects are provided so one knows when to stop taking the drug.

In the sentence: *If you feel side effects, as for example squeamishness of stomach, sweating, shaking, weakness, giddiness, dryness in the mouth, sleepiness, sleeplessness, diarrhea, reduced appetite, nervousness, excitement, headache, or sexual troubles, please ask your doctor to modify the dosage.* To find these parts in the text cam be difficult because in many cases, separators (commas) are not used, and the text continues on a new line. A pause module has been developed to handle such long lists. The special prosody module of the Profivox-Med TTS system can handle all of these problems.

During laboratory testing of this module, hundreds of PILs have been listened to and critiqued by one expert, and a phonetician constructed new rules for the TTS module based on these remarks. By relistening the same texts, fewer remarks were given by the tester. Again, new rules were developed and so on. Finally this module received 34 new rules for improving the original pausing strategy and prosody rules have been extended by 12 new rules. The use of this module has made the synthesized pharmaceutical text more understandable in the majority of cases. It must be emphasized that a completion solution to this problem is not possible until machines are capable of understanding the meaning of the text. For example, in many cases, a pause is needed before or after a reference between brackets, but in other cases, this pause is disturbing.

Text preparation for TTS conversion and database storage. Two text representation forms are utilized: chapters and sentence selection.

Chapters serve to divide the text of PILs into smaller parts. Why was it needed? The overall text of PILs is rather long, in most cases. To read this long text to the patient as one unit is not the most optimal solution. Patients may want to hear only one part of the leaflet. The NIP of Hungary also earlier suggested dividing the text of the PIL into chapters. To this end, every PIL contains 5 chapters such as *What the drug is good for? Before using. How to use? Side effects. How to store?* The TTS conversion also was organized by these chapters. So the dialogue controller offers the chapter titles for the user, and after selection, the system will read only the selected chapter. The actual selection by the user is made by voice by having the patient say the chapter title.

Sentence selection is also a special feature when preparing the texts of PILs for TTS conversion. As there may be identical sentences in the texts of all PILs, it was decided to store each of these sentences only once in the database. To create this database, the sentences of all PILs were compared and a sentence store was constructed in which only unique sentences were placed. These sentences have a special identifier and have been placed in part D of the database with their synthesized waveforms.

3. The spoken dialogue controller

The dialogue controller is perhaps the most sensitive part of a computer-based spoken dialogue system. The dialogue between man and machine should be optimal, i.e., not complicated, user friendly and clear to use. Therefore, the menu structure only contains three levels. The goal is to find the drug of interest as quickly as

possible. The dialogue begins (after the basic welcome) with the caller saying the name of the drug she/he wants to select from the database. The ASR module selects the drug, the TTS repeats the name of the drug back to the user for accuracy. Once confirmed, the user verbally selects the desired chapter of the PIL from the presented alternatives. The TTS then begins to read the chapter. The stop/continue speaking, repeat the sentence, jump functions make it easier for the user to navigate within the synthesized text of the chapter. These functions are controlled by buttons on the phone. The dialogue controller also handles the selection process when more than one item is listed under the same drug name in the database (Table 2).

Table 2. Examples of having several items under the same drug name in the database. In the case of Aspirin, there are sever items; in the case of Algopyrin, there are three versions available.

The spoken drug name	The items found in the database
ASPIRIN	ASPIRIN 100 pill ASPIRIN PLUS C effervescent tablet ASPIRIN DIREKT chewing pill ASPIRIN 500 pill ASPIRIN MIGRAIN effervescent tablet ASPIRIN PROTECT 100 mg intestine-soluble pill ASPIRIN PROTECT 300 mg intestine-soluble pill
ALGOPYRIN	ALGOPYRIN 1g/2ml injection ALGOPYRIN 500 mg pill ALGOPYRIN COMPLEX pill

For this function, the dialogue controller asks the user to verbally select his/her choice by stating the order number of the item provided by the TTS. Finally a "context sensitive help" function is available to the user at every level of the menu, to make the system more user-friendly.

4. The database

Four groups of data are stored in the database (see Figure 1).

Part A contains the drug names with their identifiers. Part B has the main marginal data of the PILs and the drugs. Part C contains the identifiers for PILs, their chapters and the sentence identifiers. Part D has the sentence store, i.e. the selected sentences in text and wave formats.

Part A consists of data for the 5000 drugs, and the dialogue controller transmits the decision of the ASR here. Drug selection is completed here. In part B, the PIL is selected and part C defines which PIL (on the identifier level) and which chapter of it will be put into the voice output of the system. Part D organizes the voice output stream on the basis of the identifiers from part C. The synthetic voice is then played by the dialogue controller. The user can ask for normal or faster speed in reading (e.g., visually impaired like for speech to be presented faster).

5. Automatic updater

The automatic updater (Block 5) is a web-based administration tool used to update the drug data twice a month. This is needed because there are continuous changes in the list (i.e., new drugs are added to the list, some of them are to be deleted). This update is done automatically. The operator is continually preparing a loader file, which contains the actual name of the drugs, and the file names of the PILs. During updating, a comparison is done. The name of a new drug found in the loader file will be put into the database and the TTS generates the pronunciation form of its name. If a new PIL is found, the updater compares the sentences of the PIL with the sentences of the sentence store. If a new sentence is found, the TTS converts it into speech. The text of this new sentence and its synthesized version will be added to the sentence store (database part D). The PIL chapter identifiers and the basic information of the PIL will be added too (parts B and C).

The loader file contains several information elements about the new or changed drugs like:

- The brand name of the drug (e.g. ALGOPYRIN).
- The full name of the drug (e.g. ALGOPYRIN 500 mg pill).
- The main active substance of the drug (e.g. Metamizole sodium).
- The identifier of the drug by NIP (e.g. OGYI-T-07845).
- The filename of the PIL (e.g. bh_0000018954_20061024094848.doc).
- The drug is sold over the counter or not.

This web based tool opportune allows the operator to edit/manage the vocabulary of the TTS and the ASR for the operator.

5 Evaluation

The evaluation of this system was executed in two phases: ASR test and system evaluation. In the ASR test, we focused on the recognition of the drug names only. In the system evaluation, TTS speech was tested along with all other components (user friendliness, speed, etc.).

The ASR test was completed with six persons (three males and three females, aged between 30 and 65). They tested the recognition of 1321 drug names by phone. The task was to call up the system, pronounce the drug name and wait for the answer, and then determine whether the recognized item was correct or the system failed. In case of a wrong recognition, the test person pronounced the drug name once more and so on, until four attempts were unsuccessful. Test results are summarized in Figure 2. In 1281 cases the recognition was successful immediately, 29 drug names were recognized after the second pronunciation, 6 after the third time and 5 of them were not recognized after four tries.

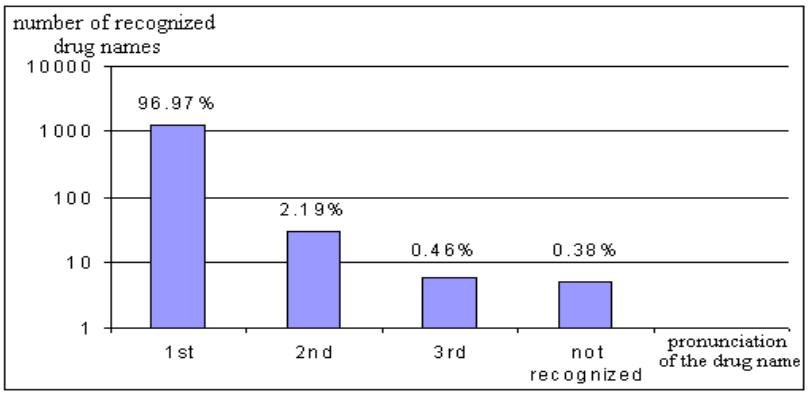


Figure 2. The recognition scale of 1321 drug names

System evaluation contained four questions: intelligibility of the voice of TTS, speed of the synthetic speech, user friendliness, and speed of getting information. Testers were asked to evaluate the system features on a scale (5 = very good, 4 = good, 3 = acceptable, 2 = less acceptable, 1 = not good. The test was completed by 57 people, 15 males and 42 females [in three age groups: 15 persons under 25 years (A), 33 between 25-60 including 7 individuals who were blind (B) and 12 over 62 years of age (C)]. Every tester got a list of 12 drug names. Their task was to call the system and listen to a minimum of two chapters about the drug. No training was given to them before testing. The results of the system evaluation are shown on Figure 3.

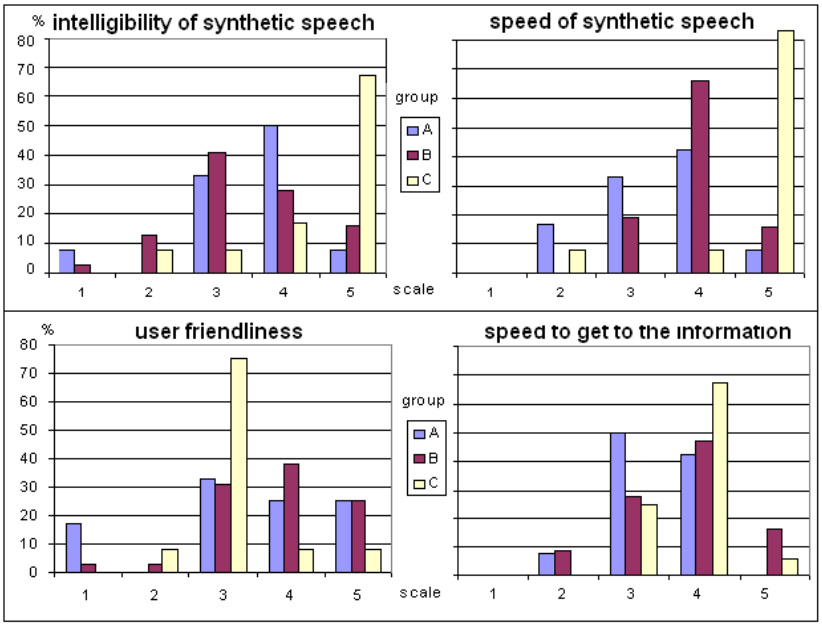


Figure 3. System evaluation results of Medicine Line

System evaluation results clearly show the difference between the young and elderly generation. Group A found the voice of TTS to be less intelligible than Group C. But as for the speed of the synthetic speech, Group C found it very good, while the younger people regarded thought it was too slow. Group C found user friendliness to be acceptable, but Group A was more positive. Only question 4 was evaluated similarly by all groups. The total system evaluation results are as follows: 1 = 2.25%; 2 = 6.25%; 3 = 30.75%; 4 = 39.25%; 5 = 21.5 %. According to these results, the system received largely acceptable and very good levels.

6 Conclusions

The experiences in using speech technology in this medical field are encouraging. The special word and expression structure of PILs required the construction of a special TTS converter and also the ASR module had to be taught in a unique way to recognize drug names. Due to ASR and synthesis features, it is clear that pharmaceutical manufacturers have to be involved into the design of such systems. As trademark laws restrict trademarks that are spelled alike, or sound similar to already existing products, the manufacturers of medications must give distinguishable names to their new products. So, perhaps linguists and phoneticians should be involved when new drug names are designed. Clearly distinguishable names can be handled by speech technology modules with a higher level of accuracy, and this is in the best interest of the manufacturers as well. The phone number of the Hungarian Medicine Line is +36-1-88-69-490. It does speak Hungarian.

7 Consortium and support

The project was realized by the Department of Telecommunications and Mediainformatics of the Budapest University of Technology and Economics, Hungary together with the National Institute of Pharmacy, Budapest, Hungary. This research and development was supported by the Hungarian National Office for Research and Technology (GVOP project no. 3.1.1-2004-05-0426), and the EU.

References

Fegyó, T., Mihajlik, P., Szarvas, M., Tatai, P. & Tatai, G. 2003. VOXenter - Intelligent voice enabled call center for Hungarian. In *Proceedings of Eurospeech-2003.* 1905-1908.

Grasso, M.A. 2003. The long term adoption of speech recognition in medical applications. In *Proceedings of the 16th IEEE Symposium on Computer-Based Medical Systems.* 257-262.

Henton, C. 2005. Bitter pills to swallow. ASR and TTS have drug problems. *International Journal of Speech Technology,* 8: 247-257.

Mihajlik, P., Tobler, Z., Tüske, Z. & Gordos, G. 2005. Evaluation and optimization of noise robust front-end technologies for the automatic recognition of Hungarian telephone speech. In *Proceedings of International Conference on Speech Communication and Technology,* Vol. 1. Lisbon, Portugal. 2677-2680.

Mohri, M., Pereira, F. & Riley, M. 2002. Weighted finite-state transducers in speech recognition. *Computer Speech and Language,* 16(1): 69-88.

Olaszy, G., Németh, G., Olaszi, P., Kiss, G., Zainkó, Cs. & Gordos, G. 2000. Profivox – a Hungarian TTS system for telecommunications applications. *International Journal of Speech Technology*, 3-4: 201-215.
Vicsi, K., Velkei, Sz., Szaszák, Gy., Borostyán, G. & Gordos G. 2006. Speech recognizer for preparing medical reports. *Infocommunication*, LXI: 14-21.
Young, S., Ollason, D., Valtchev, V. & Woodland P. 2002. *The HTK book (for HTK version 3.2.).*

Gábor Olaszy
Department of Telecommunications and Mediainformatics
Budapest University of Technology and Economics, Hungary
e-mail: olaszy@tmit.bme.hu

ON THE HUNGARIAN LANGUAGE AND SPEECH TECHNOLOGY PLATFORM

The Hungarian Language and Speech Technology Platform started its operation on June 1, 2008. The Platform was called to order by the National Office for Technology and Research. In 2007, this organization called for the formation of technology platforms in any sector of the economy. The purpose of the Call from the perspective of the Office was to launch a bottom-up catalytic process, enabling key R&D players to forge a strategic alliance, work out their strategic research agenda, and implement a plan, which would inform high level R&D policy. This was seen as a golden opportunity by the founding members of the Platform and headed by the Research Institute for Linguistics of the Hungarian Academy of Sciences. They put in a successful application.

The Platform was formed by the following partners:

Academic partners:

• Research Institute for Linguistics, Hungarian Academy of Sciences,

• Media Research Centre at the Department of Sociology and Communications, Budapest University of Technology and Economics (BME MOKK),

• Department of Telecommunications and Mediainformatics, Budapest University of Technology and Economics (BME TMIT),

• Department of Software Engineering, Szeged University.

Industrial Partners:

• Aitia Internation Inc.,

• Applied Logical Laboratory,

• Kilgray Translation Technologies,

• Morphologic Ltd.

The founding partners had been cooperating with each other in several R&D projects and have complementary expertise and resources.

Technically, the Platform is a two-year project which has the following specific objectives:

1. to develop a comprehensive Strategic Research Agenda and Implementation Plan,

2. to activate and involve all members of their field, and

3. to raise awareness of the strategic role of language and speech technology among policy makers, members of the media and the general public.

The emphasis, therefore, is not on research and development activities as such but instead on concerted actions and accompanying measures to boost the general recognition and appreciation of the field. This activity is badly needed as the field of speech and language technology is not widely known in Hungary. Speech technology colleagues often blame general ignorance or lack of interest in adopting the latest R&D achievements by key commercial players as the reason why the

Platform is needed. Much as commercial businesses care about their visual image, many companies totally neglect what may be called their acoustic image. So, the Platform's activities are aimed at bringing about a positive change of attitude, resulting in a reappraisal of the importance of language and speech technology among key industrial partners. Another important strand of activities is to reach out and involve as many industrial and commercial partners as possible.

One of the first activities of the Platform was to create a website (www.hlt-platform.hu) that not only provides information about their mission and the members of the Platform, but also provides informative pages about key areas, concepts, centres and personalities associated with speech and language technology. This website serves as a central portal for Hungarian language and speech technology. In addition, the home page of the portal displays useful demos of some of the working examples of language resources and tools, such as machine translation (www.webforditas.hu) and monolingual and bilingual parallel corpora (Hungarian National Corpus and the Hunglish Corpus respectively). The idea in showcasing these sample applications is to call attention to the fact Hungarian language and speech technology has already a lot to offer, and has a very bright future. Some of these applications are already boxed products or services, some are intriguing prototypes.

The first major event in the life of the Platform was the opening conference, which was held in Hotel Intercontinental on December 8th. The conference was designed to introduce the Platform and its partners, as well as the field of language and speech technologies. The target audience was policy makers, key industrial players, major potential users of speech and language technology, as well as the media. The whole day conference started with a plenary session in the morning, which featured two presentations, one devoted to speech technologies (speech recognition and speech synthesis) and the other to language technologies (multilinguality and semantic technologies). The European significance of the Platform was highlighted by the fact that the conference was addressed by Steven Krauwer, coordinator of the CLARIN (Common Language Resources and Technology Infrastructure) project, which is a European-wide major language technology infrastructure project. The presentations are available for download at http://www.hlt-platform.hu/konferenciahir.html.

The afternoon session was devoted to informal and interactive demo and poster sessions, in which all of the Platform members displayed their tools and resources. The conference room was literally buzzing with interest throughout the afternoon. The whole conference was recorded on video and a two disc DVD set was produced. The first DVD contains recordings of the presentations and the second contains the demo sessions. If you would like to receive a complimentary copy of the DVD set, please request them from the Platform Secretariat at info@hlt-platform.hu.

The next major effort of the Platform will focus on compiling the Strategic Research Agenda (SRA). Work is already under way to compile a survey of current activities and achievements in the field of Hungarian speech and language

technologies as preparatory background material. This will be complemented by an analysis of future trends in the field. The SRA will be ready by the summer of 2009 and will be available to the public at a subsequent conference.

Tamás Váradi
Leader of the Platform
Research Institute for Linguistics, Hungarian Academy of Sciences,
Budapest, Hungary
e-mail: varadi@nytud.hu

BOOK REVIEWS

Lingua Americana Revista de Linguistica

(Instituto de Investigaciones Literarias y Linguisticas, Universidad del Zulia, Maracaibo, Venezuela, ISSN 1316-6689)
Year 10, No. 19 (July-December 2006), 187 pp.
Year 11, No. 20 (January-June, 2007), 142 pp.
Year 12, No. 22 (January-June, 2008), 131 pp.
Year 13, No. 24 (January-June, 2009), 147 pp.

Reviewed by: **Judith Rosenhouse**
Swantech Ltd., Haifa, Israel
e-mail: swantech@013.net.il

Issues of Lingua Americana have been reviewed in this section in the past. This time, four issues from 2006 to 2009 are reviewed here. This journal, which appears twice a year, is not dedicated solely to phonetics but also to linguistic issues from South America, including General Linguistics, Applied Linguistics, and Indigenous Linguistics. The journal also has a section of book reviews. Each article has an abstract in both English and Spanish, and the articles themselves are written in one of these languages. In these issues, we find several interesting articles which are summarized here.

The 2006 issue contains seven articles in general linguistics, six articles in applied linguistics and two articles on indigenous languages. These articles deal with language learning, communication and discourse, speech/writing syntactic differences, etc. The issue ends with an index of all articles in the 2006 volume.

Among the articles in the 2006 issue, we find Elsa Mora Gallardo's "Phono-engineering at the Service of the Handicapped" (pp. 47-57), which focuses mainly on the description of the phonetic work behind a new text-to-speech Venezuelan Spanish computerized synthesizer. There are currently two computerized synthesizers being used in two centers for the handicapped – one by the educational centre (CAIDV) of Merida and the other by the faculty of humanities and education of the ULA university. Both institutes use this device with hearing impaired individuals.

The paper by Rosa Amelya Asuaje "The Rhythm of the Spanish Spoken in Venezuela" (pp. 58-66) and "Phoneme /d/ Variants in Two Different Speech Styles in Apartaderos (Meride State, Venezuela)" by Julio Marquez, Darcy Rojas and Thania Villamizar (pp. 67-73) deal with more specific phonetic issues in Venezuelan speech. The paper by Asuaje gives a general description of the role of prosody, mainly speech rhythm, in Venezuelan Spanish; however, experimental results are

not presented. The paper by Marquez et al. shows experimental results of a study of two styles of speech (formal or controlled vs. non-formal or spontaneous) which affect the articulation of the sonorous /d/. In the formal style, the stop is mainly retained, but less so (to full elimination) in the other speech style. This reduction process depends on the adjacent segmental environment (e.g., vowels vs. consonants, and which phonemes).

The volume from June 2007 contains four articles on general linguistics, four on applied linguistics and one on indigenous languages. Enrique Obediente S., the author of the first paper, entitled "/ll/ and /y/ in the Province of Merida (Venezuela) up to the Middle of the XVII Century" (pp. 9-21), deals with a subject in historical phonetics. Based on a document from that period, the study shows that there was a systematic distinction between these two graphemes and suggests that the process of "yerismo" had not yet begun at that time. This is in line with results by other researchers of American Spanish in the Indies.

Another interesting paper involving the comparison between oral speech and written expression, is "Hablaste or Hablastes: a Morphophonological Variation Case in the Spanish of Caracas" (pp. 34-53) by Giovanna D'Aquino Ruiz. This paper studies the use of 2nd person "s" in the past indicative tense in 16 speakers, evenly distributed between young (15-25 years) and older adults (50+ years) of men and women from two socio-economic (SE) levels (high and low) in Caracas. The results show that women of the lower SE level use the "s" more than high-SE level women, and more than low-SE level men. In writing, high SE level women use the "s" much more than all the other groups. The author suggests that the results reflect the mental paradigms of these speakers, in that this "s" exists in written, but not in oral Spanish.

The paper, "The Integration of Pronunciation with Other Areas of the Teaching of a Second Language" by Bertha Chela Flores (pp. 69-78), deals with the well-known topic of how to teach the phonetics of a second language. After analyzing the problem, the author suggests various methods and exercises to enhance better phonetic acquisition of the second language, including prosody (intonation rises and falls and pauses).

The articles in the third issue, from 2008, did not relate directly to phonetics, though the articles are interesting and discuss highly popular present-day topics.

The last paper to be discussed here is from the 2009 issue. The article is by Gretel Hernandez, Jorge Gonzalez and Anders Algara and is entitled "Postnuclear Alveolar Sibilant phonemes in the EFL of Upper-Intermediate Venezuelan Students" (pp. 23-42). This article deals with how university students in higher level English as foreign language courses deal with the s/z phonemes as they occur in morpho-phonological endings of nouns and verbs. The authors describe the related issues in the Spanish system and compare it with English. This issues mainly involve the different syllable coda structures, which is more variant in Spanish than in English. Differences between English and Spanish also include the fact that in Spanish, the sibilant alveolar has three phonetic allophones (/s, h, zero/) whereas English has only /s, z/. Sentences with words containing the English sibilants in various phonetic

positions were given to 14 students. The results revealed students' errors (including deletions) predominantly when /z/ was required, but /s/ was not correctly pronounced every time. The errors depended on the position of the sibilant in the word, i.e., the preceding and following consonant, vowel, or juncture. The results are discussed in light of EFL application and teaching methodology.

We have reviewed several articles from *Lingua Americana* because we do not often read about the linguistic situation in the Spanish-speaking countries in South America and the current research efforts there. The framework of the journal seems to allow rather short articles (few are more than 10 pages long). Thus, some articles are rather like research reports and do not go into details or deep discussions, while others are more general and do not report experimental results. In addition to abstracts in Spanish, there are English abstracts which are helpful for non-native readers of Spanish (who know English better). But at least one abstract (from the article "Theoretical Approximations to the Use of Unstressed Forms of the Present Pronoun in the Speech of Maracaibo" by Rosa E. Sanchez Doreste, 2008: 9-35) has appears to have been machine translated, due to various muddled structures and sentences. (The paper in Spanish shows no such problems.) In sum, the journal offers a considerable service to the linguistic study (including phonetics) of Spanish in Venezuela and other South American countries, regions in dire need of such research.

Sánchez Miret, Fernando: La diptongación en las lenguas románicas

LINCOM Studies in Romance Linguistics 04
München-Newcastle: Lincom Europa, 1998
(VI + 266 pp., ISBN 3-89586-557-5, 78 €)

Reviewed by: **José María García Martín**
Dept. of Philology, Universidad de Cádiz, Cádiz, Spain
e-mail: josemaria.garcia@uca.es

This book is a doctoral thesis supervised by the well-known Romanist Carmen Pensado at the University of Salamanca (Spain). In effect, the argument of the thesis is maintained, from start to finish, with an admirable coherence, meticulousness and a masterful handling of data and concepts. In the substantial Introduction (pp. 7-10), Sanchez Miret (henceforth SM) establishes an objective. His intention is to develop the theory of extension to its maximal potential. This theory can be used to explain developments within all Romance developments, not only those that affect the **diphthongization** (henceforth D) of /ɛ, ɔ/, but also all other vowels (p. 8). He offers a theoretical framework, Natural Phonology and Morphology, in order to provide an authentic explanation (p. 8).

Chapter I *Diptongo* (pp. 13-39) examines the nature of diphthongs from **three** points of view: phonetic, phonological and morphophonological. The conclusion of

the analysis is that diphthongs have an "ambivalent nature", since they lie in a central position between the unit (monophthong) and the duality (hiatus). This is possible because phonetically there are no reasons to consider them to be net categories, but rather it is necessary to understand them in gradual terms (cf. p. 26). Thus, the unit and the duality correspond to contrasts between one or two stable moments, prolonged or abrupt transitions and lesser or greater total durations (pp. 18 and 27). This oscillating nature produces, in the opinion of the author, a result that can, at first sight, be disconcerting: A diphthong in one language may not be considered so in another. Furthermore, the phonological analysis of the diphthongs is unique to each language (cf. p. 33). In this sense, for SM, the internal organization of the diphthong is governed by the perceptibility principle, that is, the nucleus of the diphthong is its more perceivable part, which, in general, is the most open part (p. 28). Considering this principle, the existence of decreasing diphthongs (with the more open element in the first position) and increasing diphthongs (with that element in the second position) can be justified, not by the presence of the accent in one of its elements (there are atonic diphthongs), but by its relation with syllabicity (cf. p. 34). Behind the insistence on this discrimination is the belief that overlooking this fact has hitherto made a coherent analysis of the D of /ɛ, ɜ/ in the Romance languages difficult. In spite of the attempts to establish differences between both types, we have seen that they can both undergo the same type of evolutions and the same structural solidarity that occurs between the nucleus and a post-nuclear glide or a pre-nuclear one (p. 39). The types of fundamental diphthongs, in SM's opinion, are: 1) those that oppose increasing or ascending diphthongs, in which the glide goes in front of the nucleus, to decreasing or descendent ones, where the glide goes behind the nucleus; and 2) those that separate centralizing diphthongs, in which the movement goes from out to in within the vowel space, from the peripherizing ones, which have an inverse trajectory.

Chapter 2 *Diptongación* (pp. 41-73) analyzes different processes of the D of oral vowels in diverse languages. It establishes the general principles that regulate it, which are then applied to the D of open vowels in Romance languages. In accordance with the theories of natural phonology, D is considered to be a phonological process of reinforcement in strong contexts (tonic syllable) where the vowel is longer and in slow and emphatic styles. To illustrate the author uses the description of Duraffour in his study of the Franco-Provençal dialects of Vaux:

> «A single vowel, which is not necessarily open, fragments, splits, because of the double effect of the displacements of the point of articulation and the variation (that happens sooner or later, increasing or decreasing) of the articulatory effort, into two elements, which, saving their original cohesion the most often, are liable to react to each other and to "distinguish" themselves progressively.»

Having distinguished between D and formation of a diphthong by other causes (vocalization of consonant, transformation of hiatus in diphthong, epenthesis of a

vocal sound, etc.), SM reviews the distinction between spontaneous D and conditioned D. After analyzing eight groups of examples, he concludes that the difference is not always clear, since sometimes it is difficult to know if we are in the presence of D or another process. In 2.2., *Vocales que diptongan* (pp. 49-53), the author, starts with the **assumption** that D takes place because it is impossible to maintain an **identical** vowel throughout this type of long vowel. This assumption is also based on the study of various languages which have established a hierarchy of different factors that enable vowel D to take place. Finally, this belief derives from the study of several factors in various languages; specifically, accent, intrinsic vowel duration, colour and tension. Section 2.3., *Contextos de la diptongación* (pp. 53-61), demonstrates that certain contexts can make a vowel long, such as rhythm, sentence stress, syllabic structure, the segmental context, the position of the accent and style. In section 2.4., *Resultado de la diptongación* (pp. 61-73) SM studies the D of several languages at different periods. After this analysis the author arrives at five principles: 1) D can be peripherizing and centralizing; 2) Lax vowels usually undergo centralizing D whereas tense vowels undergo peripherizing; 3) Ds are predominantly descendent; 4) Decreasing diphthongs can be transformed into increasing ones; the inverse evolution is improbable; And 5) In the process of D, a dissimilation of the characteristics of a vowel takes place, allowing some degree of prediction as to the possible results (cf. p. 62).

Chapter 3 *Introducción a los datos* (pp. 77-80) presents the results of the evolution of Latin /ĕ, ŏ/ in Romance languages [Sardinian, Provençal, Catalan, French, Italian, Castilian (= Spanish), Rumanian, Dalmatian, Portuguese, and Southern Italy dialects]. It is demonstrated that the D of these two vowels is the most widespread and it seems to have been conditioned by factors, such as the open syllable, the presence of a palatal sound or metaphony (cf. p. 80).

Chapter 4 *Teorías de la diptongación de /ɛ, ɔ/* (pp. 81-112) reviews the theories of the extension, metaphony, those of the structuralists and others which do not fit under the previous headings. None of the theories explains the Romance developments satisfactorily. Metaphonetic theory, which is correct in its pan-romance standpoint, does not rely on phonetic facts. The theory of extension provides a valid phonetic reason (a long vowel fragments itself and its components are differentiated), but almost totally links D with the open syllable and does not discuss the D before final close D vowels, nor before palatals; considering these phenomena to be of a different nature (cf. p. 110). As far as the structuralist theories are concerned, they are rejected either because of the inadequacy of the framework in which the problem resides (i.e., the loss of quantity in the Latin vocal system), or by the insufficiency of explanations, such as the preservation of distinctions or the tendency towards symmetry in the vowel systems (cf. pp. 105-106). SM also criticizes Burger's theory on D as the creation of a transition soundby considering an occurrence restricted to old French, which is not generally applicable to the other Romance languages (cf. p. 107). Spore's thesis, which first proposes a general and then restricted D, is rejected by our author, who believes the process occurs the other

way around. SM also criticizes Von Wartburg, among other authors, for resorting to external factors which imply that the causes of D are different in the diverse Romance languages (cf. pp. 107-108). I will deal with the objections to Dámaso Alonso's theories later.

Chapter 5 *Reconstrucción* (pp. 113-237) deals with the reconstruction of the processes of D in many Romance languages (Castilian, Dalmatian, Rumanian, Friulian, Italian, Southern Italy dialects, Franco-Provençal, French, Provençal, and Catalan). First, it justifies that the essential aspect of D is that the open vowels give way to centralizing diphthongs and the closed vowels to peripherizing diphthongs (p. 119). The centralizing D in principle would produce decreasing diphthongs. The successive phases of transition from decreasing to increasing diphthongs are due to successive dissimilations between nucleus and glide and the change of syllabicity in the final phase, the beginning of perceptibility (cf. p. 117). In addition, it is necessary to understand how phonological change can diversify its scope of application and it is essential to suitably interpret the contexts in which the D takes place (p. 120). SM holds that the determining factor in Romance D is the duration of the vowel, always considering the universal hierarchies of D, so that the different Romance solutions represent the varied outcomes reached within these hierarchies. SM holds that Romance languages have processes of accent reinforcement which lead to the D of the tonic vowels and the weakening of atonic ones in parallel (see parallel processes, p. 132). These hierarchies are: tonicity of the vowel, position of the accent, syllabic structure, type of vowel and sentence stress (the latter is less important). Next, the process of D is studied in those languages in which it was most intense (i.e., Castilian, Dalmatian, Rumanian, Friulian, and Italian). In these languages, with the exception of the last one, D takes place both in free and bound syllables. SM's aim here is to underline the common lines of evolution. He therefore does not go into details about the nature of diphthongs in each and every language - this information, he says, can be found in the respective historical grammars. Rather, he concentrates on those aspects which show signs of being global processes and universal hierarchies. Considering this, the conclusion is that in all cases processes of D adjust to the universal hierarchies of the process. This fact has allowed SM to reject the diverse interpretations that have tried to explain these phenomena by positing different external motivations for each language. SM also holds that unmotivated and unnecessary hypotheses about syllabification have been suggested by 19th and 20th century scholars (e.g., Ten Brink, Richter, Juret, Bourciez, Straka, Elcock - pp. 84-86 and 172).

Concerning the dialects of Southern Italy, SM demonstrates that meta-phonetic influences could not have caused phonetic D, since this would imply that the assimilation of the final atonic vowel had affected only the first part of the articulation of the tonic vowel. If this had been the case, coarticulation (= the influence of segments on adjacent others and the contextual variability which such influence produces) would have spread beyond the contiguous segments and would have acted from the beginning of the affected sound or from a certain point without

being interrupted. However, these conditions are not fulfilled in the supposed metaphonetic D of Southern Italy dialects. The author verifies that the D of /ɛ, ɔ/ in these dialects is bound in principle to those contexts of greater tonic vowel duration (before final *-i, -u*). The result of this D coincides with that of metaphony and leads to the morphologization of the alternating diphthong ~ where a monophthong marks the same grammar categories, like gender and number in nouns (cf. p. 212).

With respect to French, SM believes that the cause of the D in mid-close and open vowels in various contexts with yod and wau is not palatality or velarity, but instead the greater length of the vowels (similar to that of the free syllable) in such situations (p. 226). Solutions for French are proposed in the same way (lengthening of the vowel to be diphthongized) for the various cases of D before yod at a distance in French. SM explains D cases also in Provençal as presenting vowel reinforcement imposed with greater ease when the language preferences allow vowel lengthening, as in this language (cf. p. 234). On the other hand, he considers Catalan to be a language which is refractory to D (p. 237). SM has notably extended his considerations of Catalan in a later article (SM 2004), in which he confirms his original conclusion regarding the absence of D in Catalan. This is the only such case in Romance languages.

In the final section, the author concludes that both the D of /ɛ, ɔ/, common to many languages, and that of /e, o/ and occasionally other vowels, can be explained in the same way. It is not necessary to separate them, as Schürr and most of the Romanists subsequently did, into a conditioned D of /ɛ, ɔ/ and a spontaneous D of the rest of the vowels (cf. p. 237), as proposed by the pan-romance hypothesis defended in the conclusions of several chapters.

The book ends with an exhaustive *Bibliography* (pp. 239-260), and an *Appendix* (pp. 261-262) listing the different hierarchies of the process of D, which is very useful as a guide to help understand the argumentation. I cannot finish without mentioning that SM's text is a strong foundation for further investigation as indicated frequently within the text (pp. 152-153, 156, 165, 169, 187, 200, 219, 227 and 237).

SM's work is a magnificent, intellectual construction, which starts with an idea which I completely agree with, namely a pan-romance explanation. But I am not in total agreement with the need for a unitary principle of explanation, nor with its relevance, even if the facts facilitate it. In contrast to Dámaso Alonso's eclectic position, for example, SM proposes an explanation based on a single explanatory principle - vowel duration, which he supported by hypothesizing the generalization of changes within languages which he considered suitable (p. 109). In all aspects of human existence, actions are rarely the consequence of a single factor. Instead, diverse elements act on a single situation. Why can we not view these different hierarchies as factors stemming from multiple causations, as suggested by Malkiel (1983, Section I)? And if, as he states on different occasions throughout the work, SM believes that the Romance D is basically the result of vowel lengthening which later diphthongizes, this unique cause implies a meta-condition, a condition that is

maintained in the different manifestations that D in the different Romance languages adopts and then explains the forms they take (pp. 96-97).

This meta-condition is accent intensity (p. 128). Acceptance of this idea is difficult. SM considers the controversy over the nature of the Latin accent to have been resolved in favour of **intensity.** Nevertheless, accent does not depend on a single element (i.e., tone, intensity and duration all contribute to accent formation: p. 49). According to Quilis (1999), tone in modern Spanish is the most important of the three properties noted to explain the perception of accent. Experimental studies carried out between 1961 and 1988 indicate that the tone is the first factor in the pronunciation and perception of accent in Spanish, followed by duration and intensity, in that order (Quilis 1999: 399-400). And, finally, it is not so clear that accent in classical Latin was exclusively or predominantly related to intensity (cf. Ruiz de Elvira 1988, Lahiri/Riad/Jacobs in Van der Hulst 1999: 378-381). Clearly, based on what we know of some of the present Romance languages and what could have occurred in Latin, it is dangerous to be so categorically in favour of one of the options as we are on slippery ground. It is true that SM concentrates primarily on the time interval between Latin up to the period of Romance language formation (as declared on pp. 114, 119 and 237), and that, in the transition from classical Latin to its subsequent form, the intensive accent becomes predominant, which, without doubt, already used the popular pronunciation and sounded heavy and rustic to the refined ears of the educated (Alarcos, 1991: 212).

But we should not forget that this happens today in these languages. That is to say, returning to SM's reasoning, if the theory of lengthening in SM's version is a good test of the reinforcement of the intensity of the accent in the transition to the Romance languages, what one must demonstrate is how an accent with possible predominance, as in Latin, spawns another one of a similar type in modern Spanish after passing through an intermediate phase in which intensity is the hegemonic characteristic. Nor should we ignore that duration has a more important role in accent formation than SM recognizes, because it establishes a thread of union between the explanatory principle[1] and the metacondition[2] which would be interesting to explore. Another point that is worthy of comment is the attribution of accentual character to the rhythm of Romance languages. Perhaps it is a question of degree, but one should not forget that in comparison to English, which is a Germanic language, modern-day Spanish, a Romance language, has a syllable-timed rhythm rather than a stress-timed rhythm (cf. Stockwell & Bowen 1965: 33-34; Quilis & Fernández 1969: 160-161). In any case, Spanish seems to be a language of hard

1. There is only one explanatory principle (vowel duration) and one meta-condition (intensity accent) in SM's work.

2. See the section *Estudis linguistics*, Vol. IX in the journal *Quaderns de Filologia* (which also has the section *Estudis literaris*). This Volume IX is entitled *Lingüística diacrónica contrastiva / Lingüística diacrònica contrastive*, in Spanish and Catalan respectively. The only difference between the adjectives in this title is the accent on the second word: *diacrónica / diacrònica.*

rhythmic classification (cf. Toledo 1988: 165-168). These observations, however, do not detract from the fact that through this work, SM has decisively contributed to a better understanding of the mechanisms that have acted on the D of the Latin vowels in Romance languages and that its overall view places us in a much better position to analyse individual cases with greater accuracy.

References

Alarcos Llorach, E. 1991[4] (8th impression). *Fonología española*. Madrid: Gredos.

Hulst, H. van der (ed.) 1999. *Word prosodic systems in the languages of Europe.* Berlin-New York: Mouton de Gruyter.

Malkiel, Y. 1983. *From particular to general linguistics. Selected essays 1965-1978.* Amsterdam: John Benjamins.

Quilis, A. 1999[2]. *Tratado de fonología y fonética españolas*. Madrid: Gredos.

Quilis, A. & Fernández, J. A. 1969[4]. *Curso de fonética y fonología españolas para estudiantes angloamericanos*. Madrid: C.S.I.C.

Ruiz de Elvira y Serra, M. R. 1988. Sobre el acento latino. *Quadernos de Filología Clásica*, XXXI: 295-306.

Sánchez Miret, F. 2004. El problema de la hipotètica diftongació en català en el context de les altres llengües romàniques. In Calvo, C., Casanova, E. & Satorre, F. J. (eds.): *Quaderns de Filologia. Estudis lingüístics IX: Lingüística diacrónica contrastiva / Lingüística diacrònica contrastiva***. Valencia: Universitat de València, Facultat de Filologia. 141-188.

Stockwell, R. P. & Bowen, J. D. 1965. *The sounds of English and Spanish*. Chicago-London: The University of Chicago Press.

Toledo, G. A. 1988. *El ritmo en el español. Estudio fonético con base computacional*. Madrid: Gredos.

Daniel Jones (edited by Peter Roach, James Hartman and Jane Setter): Cambridge English Pronouncing Dictionary.

New sound recording technology supports a traditional pronunciation dictionary

Cambridge: Cambridge University Press, 2002

Reviewed by: **Wiktor Jassem**
Institute of Fundamental Technological Research, Polish Academy of Sciences, Poznań, Poland
e-mail: wjassem@amu.edu.pl

Over the last 90 years or so, at least in Europe, few reference books have been as popular among EFL (English as a Foreign Language) teachers and students as the legendary *English Pronouncing Dictionary* (EPD). Its first edition appeared in 1917 as a natural extension of Michaelis and Jones' (1913) unusual publication in which the headwords were arranged according to their IPA transcription rather than their spelling. That dictionary answered the less frequent question – *What is the spelling of the word I can transcribe phonetically as…?* The more common question is: *What is the pronunciation of the word spelled…?*. D. Jones himself edited 12 successive editions of EPD. The subsequent ones were modified and amplified by Gimson and

Ramsaran (editions 13 and 14), and by Roach, Hartman and Setter (editions 15 and 16). This 16th edition is an altogether new entity due to the inclusion of a CD.

It is virtually impossible to learn the pronunciation of a foreign language merely by reading a transcription, even if supported by extensive explanations and descriptions. Since the early 1930's, the student of 'RP', i.e. Standard British English (or BBC English, as P. Roach prefers to call it) could have, apart from a competent native teacher, audio help in the form of continuous read texts on gramophone records, and after World War II – on magnetic tape. With even a simple home computer provided with a sound card plus loudspeaker/headphones, they can now download and play back (within the usual copyright limitations) any radio theatre show broadcast by the BBC with excellent sound quality and no interference. This is very helpful, but does not relieve the student of the necessity to learn the pronunciation of individual words.

The body of the Dictionary appears on the CD with an icon next to each headword which the student can click to hear the word. They can also record themselves repeating the headword and play back their recording. They can self-judge how close their imitation was to the genuine article. But if they are prepared to use a little more effort, they can also turn both the word as spoken on the CD and as spoken by themselves into a visible picture of these speech productions. It is widely accepted that dynamic spectrograms are visual representations of the speech signal that can relatively simply be correlated with perceptually relevant and speaker-controllable phonetic features.

Visible Speech has been with us for well over half a century. Its application in teaching speech to people with significant hearing loss or total deafness was its earliest application, with controversial results. But it has never been formally and systematically tested (at least not on a convincing scale) in Foreign Language Teaching, where the students' hearing is typically normal. This situation is radically different because a normally-hearing subject is in a position to correlate visual with auditory (as well as kinesthetic and tactile) sensations needed for speech production. We would like to suggest that research in this area might be (re-)instituted now that we have easily manageable pronouncing dictionaries with recordings plus feedback. Any of the several kinds of special commercial software or a freeware speech tool, like PRAAT or Wavesurfer, which pair visual with auditory representations of speech, could be used for this purpose. The main difficulty that has to be faced is the much discussed variability-invariance issue, especially with respect to the interaction between speaker-related and language-related values of the specified acoustic features.

Six speakers, three male and three female, pronounced the headwords in (C)EPD in a random sequence. It is not always easy to determine which one of the anonymous speakers is performing since the recording conditions were not kept constant, resulting in two or more items which on close analysis turn out to be the same voice, sounding as if they were spoken by different people. Conversely, different voices were sometimes hard to keep apart: in some cases, one is at first not

quite sure whether the word is spoken by a low female or a high male voice. The significance of this imperfection to the teaching of pronunciation can be debated. At any rate, the publishers not only fail to reveal the identity of the speakers but don't even mention how many speakers are used. Figure 1 shows the mean frequencies of Formants 1 and 2 averaged separately for the male and the female voices on the accompanying CD. These results are based on measurements of vowel-targets.

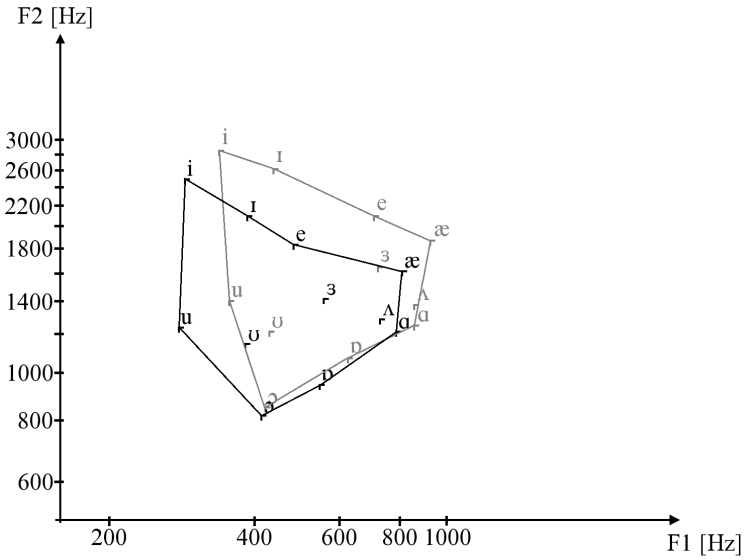


Figure 1. Target frequencies of the vowels in the (F_1, F_2) plane. Data from (C)EPD.

In view of the known correlations between the frequencies of (F_1, F_2) and the traditional description of vowels as plotted in the Cardinal-Vowels plane, our Figure confirms the agreement between the realizations of the BBC-English vowels by the speakers on the CD with their representations in Figure 1 on p. viii of (C)EPD[16]. Our results also confirm the known effect of speaker-gender on formant frequencies. The interaction between speaker-related and language-related effects does not need to be a serious problem in a systematic application of visible speech in learning the pronunciation of L_2, such as BBC English. Being offered the possibility of comparing their own version of a word (both aurally and visually) with the model on the CD, the student must answer the question, "Is this good enough?"

This is part of a wider issue: What does it mean to expect an EFL student to speak (as close as possible) without a foreign accent? Or, more narrowly, when is the student's imitation acceptable, or "correct"? And – *What is the objective criterion?* (Apart, that is, from the judgment by an expert or a native speaker.) Even if equipped with all the up-to-date paraphernalia and the ability to analyze and experiment with their productions, the student may puzzle over two spectrograms of a particular item in the Dictionary: the version on the CD and his own imitation. More specifically still: How close to the model's vowel formant frequencies should

those of the student be? The difficulty lies in the interaction of the speaker-related factors and linguistic variability.

From the earliest editions of the dictionary down to the present one, there is, in the Introduction to the Dictionary, an illustration showing the 'RP' vowels on the Cardinal Vowel quadrilateral. A comparison of the one in this 16th edition with those in the earlier editions reveals progressive changes in RP vowels over 20th-century Standard British English. These are, in a broad outline, as follows: /ʌ/ has moved forward and opened, /ɑ/ has moved upwards slightly, and /æ/ (after some vacillation) has now settled near to Cardinal [a]. The diphthongization of /iː/ and /uː/, so characteristic of Australian English, has been checked in the recent decades with fronting of /uː/. A very noticeable development is the strong raising of /ɔː/ from a half-open to a half-close position.

The arrangement of the points in our Figure shows good agreement with those on the CV quadrilateral in the Introduction to the present edition of the Dictionary on p. viii.

The question now arises how a plot like the one in Figure 1 can help the student.

The student might first learn, using real-time visible speech, to control the production of vowels, using F_1 and F_2. The next step could be to check his attempts to imitate the model using spectrograms. There is no need to try and get *the same* formant frequency values as those shown in our Figure. If the student's 'personal' English-Vowel *configurations* are *similar,* they can assume they have succeeded. It should be noted though that the /ʌ/ advocated in this 16th edition is not fronted enough and sounds rather conservative.

In order to preserve the orientation of the Cardinal Vowel quadrilateral on an (F_1, F_2) chart many phoneticians have recently used inverse scaling of both axes, with the values increasing to the left and downwards. This is quite confusing since it has for centuries been accepted, both in the natural and the social sciences, that when values are plotted in one or in two dimensions, they increase to the right and upwards. Adapting the orientation of the CV quadrilateral accordingly would imply that the talking head was assumed to be recumbent, facing up with the apex to the left, which, after all, is not unnatural. Alternatively, the acoustic chart with the normal directions of the axes could be provided with such words as "open vowels" on the right, "closed vowels" on the left, "front vowels" above and "back vowels" below. This would not be too difficult to do to insure clarity.

The CD can essentially be used in two modes. The options are termed "text search" and "sound search". The former is used to find (i.e., to hear) the selected word. The latter prints all words according to a selected sequence of phonemes. For instance, if you select /bʌ*/, you will find all words beginning with /bʌ/, and if you choose, say, /?ʌ?/ you can get all monosyllabic words that have /ʌ/ between two single consonants (like *but, cut, bug, hug,* etc.) There is a choice between UK and USA (i.e., BBC and GenAm) English, but, at least in the present edition, only the former is recorded.

The CD also contains some exercises. These are primarily intended to train the student to recognize phonemes and transcribe utterances, rather than pronouncing them. In sum, the 16th edition of (C)EPD is a milestone in the history of teaching practical English phonetics. With the present advances in speech technology, the Editors and the Publisher may be preparing other welcome surpises in the ensuing editions.

Reference

Michaelis, H. & Jones, D. 1913. *A phonetic dictionary of the English language.* Hanover: Meyer, Hachette.

Watson, Janet C.E.: The phonology and morphology of Arabic

Oxford: Oxford University Press, 2002
(Including: preface, acknowledgments, abbreviations, references, index of authors, index of subjects, hard cover, 307 pp. ISBN 0-19-925759-0, Price £75.-)

Reviewed by: **Judith Rosenhouse**
Swantech Ltd., Haifa, Israel
e-mail: swantech@013.net.il

Professor Watson, currently at Salford University, is a well known researcher of Arabic dialects. She has already published several books on the dialect of San'aa, where she spent quite some time. Her new book „The Phonology and Morphology of Arabic" (2002) seems interesting to review here. The book contains ten chapters dealing with the phonology and morphology of Arabic, specifically describing and comparing the dialects of Yemeni San'aa and Egyptian Cairo.

The choice of Cairo and San'a dialects is of course not incidental. The San'a dialect has received relatively little study in generative works on the phonology and morphology of Arabic and is closer to Classical Arabic than Cairene is (e.g., San'ani Arabic has the three consonant triads /*t, d, t*ˤ*; s, z, s*ˤ*; t̲, d̲, d̲*ˤ/ like Classical Arabic). Since San'ani Arabic is much more conservative than Cairene Arabic, it exhibits several features which are usually associated with rural or Bedouin Arabic dialects (e.g., voiced velar /*g*/ as a reflex of Classical Arabic /*q*/). Cairene Arabic, on the other hand, is an innovative urban dialect. It has lost the Classical Arabic /*q*/ and uses the glottal stop /ʔ/ instead. In addition, the emphatics are organized in a four-way distinction for two sub-sets of the coronal consonants: stops and sibilants in emphatic/non-emphatic, voiced/voiceless contrasts (/*t, d, t*ˤ*, d*ˤ*, s, z, s*ˤ*, z*ˤ/). Word stress structure in a word with more than a single binary metrical foot also differs between these two dialects. San'ani considers the final foot to be extrametrical and therefore does not take it into account in word stress assignment (e.g., *'madrasih* 'school'), whereas Cairo Arabic assigns the stress to the head of the final peripheral foot (thus: *mad'rasa* 'school').

Chapter 1 “Introduction” (pp. 1-12) presents the main features of the phonology, morphology and syntax of the Semitic language family (in Section 1.1). Section 1.2 focuses on the spread of Arabic and its development within the Central Semitic group, and specifically on its diglossic nature, (i.e., the literary vs. colloquial dichotomy). Section 1.3 outlines the structure and goals of the book.

Chapter 2, “The phoneme system of Arabic” (pp. 13-23), treats Arabic phonemes in more detail, starting with consonants in general (2.1) and then focusing on the consonantal systems in San’a (2.2) and Cairo (2.3). The vowels are discussed in section 2.4. The discussion notes the special features of many Arabic dialects in different countries, including Maltese, and certain demographic groups (Bedouins, sedentary speakers) beyond those in Cairo and San’a Arabic.

Chapter 3 “Phonological features” (pp. 24-49) describes the Arabic root, stricture, laryngeal and place/articulatory features of the language. This chapter ends with a conclusion (3.5) which sums up the basic features and processes. The starting point is, as has been already realized in previous studies, that speech sounds are not indivisible units, but are made up of phonological features and these features occur in certain orders concerning the various groups of speech sounds. She in fact adopts the approach of Clements’ (1985) concerning feature geometry, McCarthy (1988), and others in the internal structure of the speech sounds. After describing the root, stricture and laryngeal features, Watson analyzes place/articulation features which have been frequently discussed in the literature. She also explains why she considers “labial, coronal and dorsal” as articulators, while “guttural” is an articulation zone. The author also argues for a primary vs. non-primary place of articulation in this generative manner.

Chapter 4 moves on to larger units – “Syllable structure and syllabification” (pp. 50-78). This chapter follows the generative-metrical phonology methodology using the s, m, foot and heaviness syllable elements (e.g., San’a dialect allows a CVCCC syllable structure, unlike Cairene Arabic), as well as moraic segments. The analysis begins with the single word (/*bint*/ ‘girl’) and continues to lexical word combinations in various syntactic structures, such as the prefixed definite article (/*il-mudi:r*/ ‘the manager’) and suffixed bound pronouns (/*tˤard-i*/ ‘my parcel’), then leading to various phrase structures, such as */ma: ka:n-š ~ kan-š/* ‘he was not’; /*ma: libist-š*/ ‘I/you m.s. did not wear’; /*ma katab-lak-š*/ ‘he did not write to you’. (Hyphens were added by the reviewer). This then leads to the discussion of syllabification processes and morpho-phonetic processes, such as vowel reductions in closed syllables by epenthesis and syncope (in Cairo Arabic) (e.g., */tˤardi kbi:r/* ‘my parcel is big’ < *tˤardi kibi:r*). This analysis brings the differences between San’ani and Cairene syllabification systems and phonological processes to light. In addition, she refutes Broselow’s (1976) hypothesis that syncope applies blindly in Arabic dialects to destroy vulnerable monomoraic syllables. In San’ani Arabic, this assumption is not correct because syncope applies optionally at the beginning, but not the end of the phonological word. The syncope there is partly lexical and deletes vowels in bimoraic, as well as monomoraic syllables.

Chapter 5 is on “Word stress” (pp. 79-121). This long chapter also deals with the theoretical model that can be applied to Cairene Arabic (Hayes 1995). As she explains, the possible theoretical contrast revolves around the moraic model (in which short vowels are assigned one mora, long vowels two moras and geminate consonants get one mora in the underlying structure) and the iambic/trochaic division (iambic feet are formed from two elements which contrast in length, while trochaic feet are formed which contrast in intensity). After the analysis of many examples in both dialects, the author concludes that these dialects have moraic trochee systems. Cairene lacks foot extrametricality and has special rules for certain verb forms, which are lexicalized according to the author, who assumes them to be residues of an older stage of development of this dialect. San’ani has a somewhat different syllable structure, and different rules, including stress fluctuation, especially in connected speech, and word-internal CVV and CVG (Consonant-Vowel-Geminate) syllables.

Chapter 6 moves on to discuss “Morphology” (pp. 122-174), mainly morpheme structures, root-and-pattern and non-concatenative morphology, prosodic morphology, and level-one verbal and nominal morphology. This two-level model is due to Watson’s opinion that the Arabic system should be analyzed at two morphological levels: level one covers non-concatenative processes and level two includes concatenative processes. The basic forms and patterns of the verb system are analyzed as syllable structures (mainly moraic). In the nominal morphology part, the author refers primarily to the formation of broken plurals in the two dialects (following McCarthy and Prince 1990) as well as to the main verb forms (sometimes named “measures”).

Chapter 7, “Morphology 2” (pp. 175-199), continues this discussion with what the author calls “level two” verbal and nominal morphology, mostly in Cairene Arabic. In this “level two”, she includes wholly concatenative structures of suffix morphemes in verbal and nominal conjugations and inflections. The suffix /–*i*/ is noted because it occurs in some cases (with several roles) on level 1 and in others on level 2 (cf. *masʕri* ‘Egyptian’ with the suffix following the singular form, *ganayni* ‘gardener’, with the suffix following a noun in the broken plural form, *saʕa:ti* ‘watchmaker/repairer’ where the suffix follows the pl. f. suffix). A very short section refers to special forms in San’ani Arabic because it has fewer suffix morphemes than Cairene Arabic, especially ones that are due to foreign influence, like Turkish. The metropolitan Cairene history is the reason for the difference between the two dialects in this respect.

Chapter 8 “Lexical phonology” (pp. 200-225) returns to a discussion of general phonological processes by examining the prosodic and melodic processes of various word structures in the two dialects. Among the described processes, the author notes diphthong reduction and *n* strengthening, as well as *h* disassociation in San’ani Arabic (e.g., *ʔabsarayn + ha:* → *ʔabsarannha:* ‘they f. saw her’, *gad + hu:* → *gadu:* ‘he is’, *man + hum* → *manum* ‘who are they m.?’, but: *inn ha:na:* ‘that here’ without /*h*/ reduction). Section 8.2, “Melodic Processes”, considers total assimilation

using Mohanan's (1993) approach. Watson discusses the rules which control *l*-assimilation (the definite article) when prefixed to nouns with various different consonants at their heads and the *t*- of the detransitivizing prefix (e.g., *tgawwiz* → *iggawwiz, itgawwiz* 'to get married').

Chapter 9 "Post-lexical phonology" (pp. 226-267) continues the analysis of prosodic and melodic processes in various word categories in these dialects. The sections in this chapter deal with unstressed long vowel shortening *be:t + e:n* → *bite:n* '2 houses'), vowel deletion (*bi-aktib* → *baktib*), glide formation (*iktibu + intu* → *iktibu [w]intu* 'you pl. write!'), glottal stop epenthesis (*gadu: ʔaħmar* → *gadu: aħmar* 'it is red'), and gemination of clitic-final sonorant in San'ani Arabic (e.g., *al + ism* → *all-ism* 'the name'; *min + ams* → *minn ams* 'from yesterday'). Other "melodic" processes discussed here are: nasal place assimilation; coronal sonorant assimilation; coronal place assimilation; voicing assimilation; voicing, devoicing and geminate devoicing in San'a Arabic; palatalization; labialization of labial and dorsal consonants; and labialization of dorsal vowels. These processes were described in terms of C/C, C/V and V/C interactions of two conflicting processes in the two dialects and their different results.

Chapter 10, "Emphasis" (pp. 268-287), the last chapter of the book, is devoted to this important and frequently discussed feature of the Arabic language. The term "emphasis" refers to the spread of pharyngealization from pharyngeals to vowels and consonants in the local (within the syllable) and long (beyond the syllable) distances. These pharyngeals are noted in the Central Semitic languages, while in Arabic grammatical literature, they were considered distinct from their plain counterparts. Recent laboratory works have indicated that their production involves constriction of the upper pharynx. Watson considers emphasis as pharyngealization and represents it as non-primary [guttural]. Acoustically, emphatics involve a lowered F_2 due to an enlarged mouth cavity, and raising of F_1 due to a reduced laryngeal cavity. Emphasis spread varies among dialects and most of this chapter described the various patterns of spread in the Cairene and San'ani dialects. She finds that the unmarked direction of spread of [guttural] is right to left, while the unmarked direction of [labial] spread is left to right.

In sum, the book supplies a detailed phonological outline of the structure of two Arabic dialects using modern analysis methods and theoretical models. The description involves not only segments, but also the morphological and syntactic elements which form the phonological units. This book is warmly recommended for Arabists and Semitists, but also for any reader – student or professional researcher alike – who is interested in phonological processes and phenomena from the general perspective of linguistic structure.

References

Broselow, E. 1976. *The phonology of Egyptian Arabic*. Ph.D. dissertation. University of Massachusetts.

Clements, G. N. 1985. The geometry of phonological features. In Ewen, C. & Anderson, J. (eds.): *Phonology Yearbook 11*. Cambridge: Cambridge University Press. 225-252.

Hayes, B. 1995. *Metrical stress theory: Principles and case studies*. Chicago: University of Chicago Press.
McCarthy, J. J. 1988. Feature geometry and dependency: a review. *Phonetica*, 43: 84-108.
McCarthy, J. J. & Prince, A. 1990. Prosodic morphology and templatic morphology. In Eid, M. & McCArthy, J. J. (eds.): *Perspectives on Arabic linguistics 2*. Amsterdam, Philadelphia: John Benjamins. 1-54.
Mohanan, K. P. 1993. Fields of attraction. In Goldsmith, J. (ed.): *The last phonological rule: Reflections on constraints and deviations.* Chicago: University of Chicago Press. 61-116.

Richard V. Teschner & M. Stanley Whitley: Pronouncing English. A stress-based approach with CD-ROM.

Washington: Georgtown University Press, 2004

Reviewed by: **Wiktor Jassem**
Institute of Fundamental Technological Research, Polish Academy of Sciences, Poznań, Poland
e-mail: wjassem@amu.edu.pl

This text differs, in several fundamental respects, from the traditional model of a pronunciation manual:

1. The material used in the entire book is based on a database of approx. 25,000 lexical items assumed by the authors to be representative in terms of occurrence frequency.

2. Prosody precedes segmentals in both the text itself, and in methodological conception.

3. The rhythm of GenAm English is described in terms of the units of Classical prosody, i.e. as representations of spondees, trochees, iambs, dactyls or anapests.

4. Phonetic transcription (IPA) is used only when unavoidable. The examples cited throughout the book are – in principle – only given in orthography.

5. Use of examples is very extensive (due to special computer programs) as it draws on the authors' database.

The structure of the book is as follows: Chapter 1. The metric foot; Chapter 2. Strong stresses and weak: How to know where they go; Chapter 3. Intonation - The melodic line; Chapter 4. From orthography to pronunciation; Chapter 5. Vowels; Chapter 6. Consonants; Chapter 7. Sounds and forms that change and merge; and Chapter 8. Appendix.

The authors' general ideas may be formulated like this:

1. Assuming you *know* what the sounds (the phones) of GenAm are, and how to produce and concatenate them, you learn how to use that knowledge when you are confronted with orthographic representations of words in a (connected) text.

2. The primary information you have to possess is the *location* of *lexical stress.* This information must be used by your mental engine, which combines it with rules + exceptions, converting orthography into a spoken realization of successive

phonemes, complemented by a book of 'special cases', including grapheme(s)-to-phoneme conversions of specific individual lexical items.

Chapter 1 is the most original, as it shows in detail how individual English words in their citation forms, as well as complete phrases, can be interpreted in terms of Greek/Latin prosody. While trying to be consistent, the authors find problems in their analyses, which they overcome by introducing an added principle, borrowed from certain varieties of generative phonology, called extrametricality. Extrametricality sometimes tends to brutally explain away anything that doesn't fit a preconceived model. It can "Exclude a designated peripheral element from the computations in a metrical grid" (Roca & Johnson 1999, p. 689). This definition is limited to the extent that it allows the extrametrical part of a given unit only to appear in initial or final position. But the temptation arises to squeeze any inconvenient case into the closest-fitting mold and place the overflow either at the front or the back end, preferably the latter. The reader is not informed what advantage is gained by accepting this representation, as opposed to the traditional ones.

Section 1 of Chapter 1 naturally opens with *The notion of stress: Present stress and absent/null stress.* We read: "Strong stress can be defined simplistically as the greater prominence or loudness that a vowel or syllable exhibits within a word" (p. 1). Considering that (as indicated in the book's title) the entire text is founded on the notion of (word) stress, one may be surprised to learn that the authors should be content with a "simplistic" definition. More information is given some twenty lines later: "A strong-stressed vowel […] is **louder** in volume and **longer** in duration/length and/or **differently pitched** […] than the rest of the vowels in the word" (ibid, authors' bolding). Here, the authors are in agreement with the tradition, but are at odds with much of the extensive experimental and theoretical work on the phonetic nature and function of stress (or accent) performed in the second half of the 20th century, as reviewed, e.g. by Fox (2000). We do not wish to re-open the debate here, but would suggest that a simple acoustic analysis of a phrase like *You'd better* spoken as a warning (with a final fall-rise) be made. It will show that the final **un**stressed (unaccented) /ə/ is in fact louder, much longer than the stressed /e/ and quite "differently pitched" from it. Similar examples compromising the above definition can easily be cited. All the same, this turns out to be of no consequence because what matters throughout the book is **not** the **nature** of stress, but its **location**. This premise is discussed and richly illustrated (in Chapter 2) from the following points of view: (a) part of speech, (b) effect of prefixation, (c) effect of suffixation, and (d) accentual shift within morphological families. What accentual regularity there is in the English lexicon is known to depend largely on suffixation. This aspect, like some others considered in this book, has been exhaustively discussed in Kingdon (1958a) and Fudge (1984). But neither of these monographs is mentioned either in the text or in the bibliography. Intonation is presented, in Chapter 3, with comparisons to works, such as Pike (1945) or Kingdon (1958b), in a much simplified form, and not necessarily always correctly. For instance, in part 2 of Figure 3e, the last syllable is indicated as higher than the preceding one, although

what was intended was neither “progression” or “interrogation”. On the other hand, this chapter contains observations and materials relating to the stressing of compounds (whether indicated as such in the spelling or not) which is virtually absent from other treatments of English phonetics. The notorious complexity of the relations between spelling and pronunciation in English has been the subject of extensive studies for about four centuries now, beginning with the “Orthoepists”. It is treated by Teschner and Whitley, in Chapter 4, with unrivalled thoroughness. Throughout the book all points are substantiated withappropriate exercises. Chapters 5, 6, 7 and 8, which together form a kind of outline of English phonetics, do not supply information that could not be found in other sources. Although the phonemic principle is observed, the term *(speech) sound* is not quite consistently used and refers sometimes to a phoneme, and at other times to an (allo-)phone. The most striking regional variations within American English and between BBC English (RP) and GenAm are mentioned wherever appropriate, but not discussed to any significant extent.

The accompanying CD will be of interest not only to students learning English at various stages, but also to specialists applying statistics to problems of phonetics and grapheme-phoneme relations. It contains two types of files. The text files include complete (in the sense of the authors’ Database) lists of lexical items illustrating all key points, like all words containing the (realization of the) individual phonemes, or all words stressed on the penult etc. The audio tracks are mainly intended as models to be imitated by the student.

The book is not free from editorial oversights. For instance, in Figure 4b on p. 99, most phonetic symbols are simply missing, and the slashes appear with nothing in between. On p. 133, line 1 speaks of «the digraph [read: grapheme] ‘u’». Imperfections such as those shown above can easily be spotted and, we hope, corrected in a second edition. A judicious teacher will value the book under review even now, especially for its wealth of practical materials.

References

Fox, A. 2000. *Prosodic features and prosodic structure.* Oxford: Oxford University Press.
Fudge, E. 1984. *English word-stress.* London: Allen & Unwin.
Kingdon, R. 1958a. *The groundwork of English stress.* London: Longmans, Green & Co.
Kingdon, R. 1958b. *The groundwork of English intonation.* London: Longmans, Green & Co.
Pike, K. 1945. *The intonation of American English.* The University of Michigan Press.
Roca, I. & Johnson W. 1999. *A course in phonology.* Oxford: Blackwell Publishers.

Daniel Schreier: Consonant change in English worldwide. Synchrony meets diachrony

Basingstoke, UK, and New York, USA: Palgrave Macmillan, 2005
(248 pp., inc. list of tables, and list of figures, Acknowledgments,
list of abbreviations, pp. ix-xvi, end notes pp. 226-229, references pp. 230-243,
and index pp. 244-248, Price: $85.00 Hard cover. ISBN 1-4039-9824-8)

Reviewed by: **Judith Rosenhouse**
Swantech Ltd., Haifa, Israel
e-mail: swantech@013.net.il

This book, which appeared in the Palgrave Studies in Language History and Language Change series, focuses on a particular issue: consonant change. Dr. Daniel Schreier, who has published previously on related issues (Schreier & Lavarello-Schreiuer 2003) describes and analyzes this subject in the English language, both in its original country and worldwide, starting from its past roots in Medieval English in Britain and ending in its present-day state. The book has six chapters: 1. Introduction (pp. 1-15), 2. Consonant clusters: General observations (pp.16-55); 3. Initial cluster reduction in English (pp. 56-125); 4. Final cluster reduction in English (p. 126-197); 5. Theoretical implications (pp. 198-220); and finally, 6. Summary and Conclusions (pp. 221-225). The book is based on a thorough study of the historical roots of consonant clusters in English and their variants in previous colonies and elsewhere during the period of almost one thousand years. It also describes the phonetic processes consonant clusters have undergone.

In the first chapter, the author presents an interdisciplinary approach to phonotactic language change, on which this study is based. He then discusses the relationships among typology, language universals and phonological theory, combining them using the processes of consonant deletion and insertion, which are apparently universal. The next section in this chapter refers to the notions of language variation and change, contact linguistics, genetic linguistics, and language acquisition, learning and psycholinguistics. He further notes that this book addresses issues that have received little attention in the literature on English historical linguistics. We learn that some of the historical phonotactic changes in English include the initial clusters /*kn*/ (*knee*), /*gn*/ (*gnat*), /*wr*/ (*write*), the intermediate clusters /*st*/ (*listen*) and /*ft*/ (*often*), and the final /*lx*/ (*wealh*) and /*çt*/ (*bright*) clusters. By linking past and present in this corpus-based work, this study represents an in-depth analysis of when and why consonants have changed in English and also deals with a number of questions which are still open.

Chapter 2 is the longest chapter of the book and it presents the major research questions of the whole study: 1. What consonant cluster (CC) types are found in the phonotactic system of English? 2. How are CCs modified? 3. How can one trace the diachronic evolution of English CCs? 4. What external and language internal effects

underlie this process? 5. Who are the principal agents of change? This is a rather theory-directed chapter as its section titles imply: 2.1. Consonant clusters and syllable structure; 2.2. Change and adaptation mechanisms; 2.3. Causes and motivations: a first approach; and 2.4. Conclusion.

The following study of initial and final cluster reduction in chapter 3 and 4 arrives at interesting findings. The approach taken claims that initial cluster reduction in British English did not occur all at once – it advanced gradually, beginning in the 12th century, and was consonant-dependent. The first initial cluster to reduce was *h*-initial clusters. Innovative spellings of *n-*, *l-*, *r-* clusters appeared, however, as early as the 9th century. Some of the clusters did not change until early modern English.

The present work is based on the lexical corpora found in the huge Helsinki Corpus (Kyto 1993), the Oxford English Dictionary (Murray 1878-1910) and the Anglo-Saxon Dictionary (Toller 1898). As part of this work, the author presents (pp. 68-69) three figures which depict the variations in the spelling conventions of *hr- > r-*, *hn- > n-* and *hl- > l-* in ten language periods [3 Old English, 4 Middle English and 3 (Modern) English]. Other tables and figures depict processes involving other clusters. Figure 3.5 (p. 78) summarizes the diachronic dimension of initial cluster loss between the years 850-c.1800. Following this part of the chapter, the author focuses on the reduction of *h-* clusters in various countries outside England. Section 3.2 deals with *hw > w* in New Zealand, by studying the origins of the population there (Australia, England, Ireland, Scotland and some other countries). Also in New Zealand, the study is based on corroborating evidence from spellings in numerous references. The details are listed in Table 3.7 which summarizes the maintenance and loss of *hw-* in New Zealand English by sex, region, birth date, word type, and preceding and following phonetic environment (p. 101). Several additional bar graphs describe other details of the involved processes. Initial cluster reduction [as in */sp/ (speak) > p (piki), st (stand) > t (tan), str (strong) > tr (tranga)*] is examined in this chapter using the English-based Creole of Suriname. Thus, the chapter analyzes not only language contact, but also creolization.

Final cluster reduction in English is the topic of chapter 4, which addresses the complex nature of context-sensitivity and language internal constraints, as well as the interplay of external histories and internal developments of CC reduction. In this chapter, Professor Schreier revisits the process of cluster final stop deletion and goes into typological, structural and methodological criteria and the effect of environment and resyllabification processes on final clusters. Next, he analyzes CC reduction in four varieties of English in New Zealand and the South Atlantic: St. Helena, Tristan da Cunha, Maori and Pakeha. The next section in this chapter analyzes and compares other varieties of English, including several British and American English dialects, the creolized English of Tristan de Cunha and St. Helena as well as Bahamian and Jamaican English (Creole languages), Korean English (where English is a foreign language) and Indian English (where English is institutionalized). This chapter reveals both conditions and constraints on the reduction of final CC. CC reduction is in competition with various other strategies employed (in bilingual,

koineization and other kinds of language contact) to break up unfamiliar clusters and modify unusual or unfamiliar phonotactic properties in the target language. In fact, according to the author, this chapter provides the most complete overview of word final CC reduction in the world.

The fifth chapter, Theoretical implications, provides, as a result of the previous study, several principles that operate in CC reduction. Principle 1 claims that "no speaker of English […] fully realizes all clusters in all linguistic environments and in all social contexts or speech styles" (p. 199). Thus, consonant reduction processes are universal. Principle 2 (p. 200) says that contact between systems with similar or identical phonotactic systems does not lead to phonotactic simplification. Principle 3 (p. 201) states that phonotactic transfer and change is most likely to occur in contact between language varieties with distinct phonotactic systems. Principle 4 (p. 202) says that phonotactic transfer may be short or long-term depending on various factors. Principle 5 (p. 212) adds that internal constraints on CC reduction (phonetic, morphological and cluster-typological) are as diagnostic and genetically indicative as global CC reduction values. The author also offers implications for language change, and here principle 6 (p. 216) is that cluster reduction is both an internal and an external process. Whereas, internal change typically operates on initial clusters, final CC reduction is ubiquitous and very prone to external change. A psycholinguistic explanation of the processes leads to principle 7 (p. 220): Lexical processing is a crucial factor to explain when describing why initial clusters are more stable than final ones: information lost at the beginning of words impedes word recognition, whereas information lost at the end of words often occurs at little cost since word recognition is already completed.

The summary and conclusion in Chapter 6 bring together the main themes of this study. The most frequent process is that of final CC reduction. By comparing of the varieties of English, the author concludes that English varieties worldwide are not as homogeneously patterned as often assumed. The phonetic environment of the cluster is very important and its effects are especially susceptible to the segment preceding the cluster and the morphemic status of the CC-final plosive. Differences were found as to the phonetic nature of the clusters that could undergo reduction. Schreier also outlines a putative trajectory of phonotactic developments upon language contact. "The type of strategy adopted in the first stages is variety-specific. Although there is a strong trend to insert epenthetic vowels as cluster reduction, cluster reduction is most pertinent when the substrates have canonical CV syllable structure. Then however, the production of clusters typically increases and native-speaker constraint types begin to emerge as accommodation to the target intensifies" (p. 224).

To sum up our survey of this book, we can recommend it to anyone interested in English and general phonetics. The book offers a general view of a phonetic phenomenon or process in English, now probably the most significant world language as its modern lingua franca. The methodology is interdisciplinary and combines variationist linguistics and historical change mechanisms with psycholinguistics, sociolinguistics, contact linguistics and phonological theories. As

such, it is strongly based on the literature (see the 13 page long Reference list), as well as on the author's own investigations. The style is easy and flowing so even undergraduate students could easily follow it (although undergraduates do not usually go into such deep analyses). For professionals in any of the above areas of linguistic study, this book is a real contribution.

References

Kyto, M. 1993. *Manual to the diachronic part of the Helsinki Corpus of English Texts*. 2nd edn. Helsinki: Helsinki University Printing House.

Murray, J. 1878-1910. *The Oxford English dictionary (A new English dictionary on historical principles)*. Oxford: Oxford University Press.

Schreier, D. 2003. *Isolation and language change: Sociohistorical and contemporary evidence from Tristan da Cunha English*. Houndmills/Basingstoke and New York: Palgrave Macmillan.

Schreier, D. and Lavarello-Schreier, K. 2003. *Tristan da Cunha: History, people, language*. London: Battlebridge Publications.

Toller, J. 1898. *An Anglo-Saxon dictionary*. Oxford: Clarendon Press.

Jeroen van de Weijer, Kensuke Nanjo, Tetsuo Nishihara (eds.): Voicing in Japanese

Berlin, New York: Mouton de Gruyter, 2005
(viii + 314 pp, including: Preface, Contents with List of Contributors, Bibliographical references and Index, ISBN-13: 978-3-11-018600-0, $82 US)

Reviewed by: **Eric Rosen**
University of British Columbia, Vancouver, Canada
e-mail: errosen@interchange.ubc.ca

Recent studies of Japanese phonology have been important in developing and testing proposals of current phonological theory – for example, licensing and underspecification (Itô, Mester and Padgett 1995), moraic theory (Itô and Mester 1995, Kubozono 1999), OCP effects (blocking of rendaku voicing by "Lyman's Law"), and models of lexical stratification (Itô and Mester 1995). The process of rendaku voicing (McCawley 1968; Vance 1980; Itô and Mester, 1986; Mester and Itô 1989) voices the initial obstruent of the second member of a Japanese compound word and applies mainly to compounds of Yamato (native Japanese) origin. Therefore, a volume of phonological and phonetic studies of voicing in Japanese, and mainly on rendaku, is of likely interest to a wide range of phonologists and phoneticians.

The book is organized in two parts. "Part I – Consonant Voice" contains ten papers, as follows: Haruo Kubozono "*Rendaku*: Its domain and linguistic conditions" (pp. 5-24), Keren Rice "Sequential voicing, postnasal voicing, and Lyman's Law revisited" (pp. 25-45), Kazutoshi Ohno "*Sei-daku*: diachronic developments in the writing system" (pp. 47-69), Kuniya Nasukawa "The representation of laryngeal source-

contrasts in Japanese" (pp. 71-87), Timothy J. Vance "Rendaku in inflected words" (pp. 89-103), Haruka Fukazawa and Mafuyu Kitahara "Ranking paradoxes in consonant voicing in Japanese" (pp. 105-121), Noriko Yamane-Tanaka "The implicational distribution of prenasalized stops in Japanese" (pp. 123-156), Hideki Zamma "The correlation between accentuation and Rendaku in Japanese surnames: a morphological account" (pp. 157-176), Tomoaki Takayama "A survey of rendaku in loanwords" (pp. 177-190), and Keiichiro Suzuki "Recognizing Japanese numeral-classifier combinations" (pp. 191-204).

"Part II – Vowel voice" contains the following four papers: Kikuo Maekawa and Hideaki Kikuchi "Corpus-based analysis of vowel devoicing in spontaneous Japanese: An interim report" (pp. 205-228), Mariko Kondo "Syllable structure and its acoustic effects on vowels in devoicing environments" (pp. 229-245), Miyoko Sugito "The effect of speech rate on devoiced accented vowels in Osaka Japanese" (pp. 247-260) and Shin-ichi Tanaka "Where voicing and accent meet: Their function, interaction, and opacity problems in phonological prominence" (pp. 261-276).

Although most of the main topics are familiar, many contributions offer new or little-known data that point to interesting analytical problems. Processes, such as rendaku voicing, the interaction between nasality and voicing (post-nasal voicing, prenasalization of stops, and nasalization of velars), and vowel devoicing, are discussed in a number of articles. Some articles focus mainly on presenting data and identifying the questions raised by the data, (e.g. Kubozono pp. 5-24). Others make theoretical proposals (e.g., Tanaka pp. 261-276).

Kubozono (pp. 5-24) offers lesser-known data on rendaku voicing and establishes its relevance to current theory and past analyses. He examines three limitations on rendaku voicing: (a) transmorphemic Obligatory Contour Principle (OCP) blocking of rendaku in surname N-N compounds, (b) morphological structure conditions on voicing, and (c) a prosodic condition where a mono- or bi-morphemic N1 blocks voicing in compounds with bimoraic N2 *hon* 'book'. For the (b) case, he shows that similar morphological conditions hold for other phonological processes in Japanese, English, and Chinese. For the (c) case, he gives independent evidence for a 2μ + 2μ domain: (i) contraction and a /h/ ~ /p/ alternation in Sino-Japanese bimorphemic 2μ + 2μ compounds[3], (ii) accent domain in reduplicated mimetics and (iii) accentuation of numeral sequences forming 2μ + 2μ accent-domain units.

In her article (pp. 25-45), Rice proposes a "dual mechanism hypothesis" (henceforth "DMH") in which laryngeal voicing ("LV") and sonorant voicing ("SV") are phonologically distinct.[4] The DMH accounts for the paradoxical nature of nasals, which are transparent to Lyman's Law OCP effects, yet trigger voicing in

3. Contraction and /h/ ~ /p/ support a moraic analysis less convincingly. Contraction failing to occur in > 4μ [dai-butu]-si (*dai-bus-si) or /p/ failing to surface in > 4μ [man-nen]-hitu (*man-nem-pitu) could equally be explained by morphological structure since constituent sisterhood correlates exactly with the application of the process in question.
4. See also Ohno (this volume).

following obstruents.[5] Rice's evidence for the DMH comes from the interaction between OCP blocking of rendaku and postnasal voicing. She argues that post-nasal voicing in verb-verb compounds (with no [+voi] junctural morpheme) is not blocked by the OCP in *hun-zibaru* ('tie-up') because of the non-interaction between SV in the post-nasal /z/ and LV in the voiced /b/.[6] Rice also argues that postnasal voicing is contrastive morpheme-internally, relying on (a) the nonstandard assertion that Sino-Japanese compounds such as *sam-po* 'walk' and *ken-ka* 'quarrel' are monomorphemic,[7] and (b) a small number of monomorphemic words with NT sequences that Itô and Mester (1999) consider peripheral in the lexicon. Finally, she argues against the stratification of the Japanese lexicon proposed by Itô, Mester and Padgett (1999). Her comments are a continuation of an ongoing debate between her and Itô, Mester and Padgett on this topic. (See Itô, Mester and Padgett 1999.)

Ohno (pp. 47-69) discusses the history of orthographic representation of voicing contrasts in Japanese, traditionally known as the *sei-daku* (lit. 'clear-muddy') distinction, which went through three historical stages: (a) distinction through different Chinese characters; (b) no distinction, and (c) distinction by diacritics. Ohno examines whether a phonemic voicing contrast existed at all stages and how nasalization of obstruents fit into phonemic distinctions. He looks at how orthographic evidence bears on these questions.

Nasukawa (pp. 71-87) examines the phonetic nature of voicing contrasts in Japanese in the framework of Element Theory (Kaye, Lowenstamm and Vergnaud 1990). He argues that the terms "voiced" and "voiceless" do not precisely characterize the laryngeal nature of obstruents. Instead, the monovalent "elements" [L] (negative VOT) and [H] (positive VOT) can, in various combinations, distinguish possible laryngeal source contrasts. He argues that Japanese allows only two specifications with [L] and [H]: (a) neither of [L] or [H] (voiceless unaspirated or zero VOT) and (b) [L] (truly voiced, or negative VOT). [H] alone does not occur. His proposal predicts that voiceless obstruents with no elemental specification should not trigger any processes. He therefore explains vowel devoicing by positing a "noise element" [h], in voiceless obstruents that has an acoustic rather than phonological effect on high vowels.

Vance (pp. 89-103) shows, with data from Okumura (1955), that rendaku generally does not occur in V-V (verb + verb) compounds. Historically, rendaku in N-N (noun-noun) compounds occurred as reflex of the particle *no* which could occur with a N but not a V. Apparent V-V compounds that undergo rendaku are generally

5. This is an alternative to the licensing account of Itô, Mester and Padgett 1995. Rice does not reject their account explicitly, but refers to Pater 1999, who argues against NC licensing for reasons not discussed by Rice.

6. Her case would be strengthened by offering more than one such example.

7. Rice bases this claim on supposed evidence in Vance (1996). Vance points out that many Sino-Japanese "binoms" are less transparently compositional than Yamato compounds, but nowhere does he offer more substantive evidence against the polymorphemic status of these words.

verb plus deverbal noun (henceforth V-V). Vance systematically surveys dictionaries to count rendaku occurrence in V-V and V-V compounds. He finds that a few V-V compounds do not show rendaku where expected, and a few V-V compounds do show rendaku where not expected. He also finds that compounds formed from adjectives and verbs generally undergo rendaku more often than V-V compounds. These facts are intriguing, given that phrasal combinations of verbs and adjectives would not syntactically include the particle *no*.

Fukazawa and Kitahara (pp. 105-121 and henceforth "F&K") use data from Tateishi (2001) on baseball team names borrowed from English to show that the /-z/ plural allomorph of English surfaces in Japanese as /-s/ when preceded by a voiced obstruent (e.g. *kabusu* 'Cubs'); however when the /z/ occurs post-nasally, it surfaces as voiced, even if preceded by another voiced obstruent (e.g. *iNdiaNzu* 'Indians'). Their data suggest that in the framework of Optimality Theory (henceforth "OT": Prince and Smolensky 1993), a constraint against post-nasal voiceless obstruents (*NT) must outrank a constraint against two voiced obstruents in a morpheme (*DD) contradicting the ∗DD >> ∗NT ranking proposed by Itô and Mester (1999). The authors' solution to this problem is somewhat controversial, positing that speakers of Japanese are aware of English plural morphology and analyze a word like *kabusu* as containing a plural suffix. Nevertheless, both their data and their argumentation raise interesting questions for further research.

Yamane-Tanaka (pp. 123-156) uses evidence from both synchronic variation and diachronic change among prenasalized stops in Japanese dialects to argue in the OT framework for (a) fixed harmonic scales of markedness, and (b) restrictions on constraint re-ranking, where demotion of Faithfulness constraints is the unmarked grammatical change. Prenasalization of intervocalic stops fall into the following four main dialectal patterns: {^{m}b, ^{n}d, N}, {^{n}d, N}, {N}, and {ϕ}, where the sets of stops are in successive subset relations. In addition, prenasalized stops were lost in the diachronic progression: [^{m}b], then [^{n}d], then [^{n}g]. Fixed harmonic scales proposed by Prince and Smolensky (1993) where COR > LAB (coronals are more harmonic than labials) and SIMPLEX > COMPLEX (simplex segments are more harmonic than complex ones) yields the ranking *^{m}b >> *^{n}d >> *N. When Faithfulness (F) interacts with this fixed ranking, only four rankings are possible. The rankings predict attested inventories of prenasalized stops across dialects. Demotion of Faithfulness captures the attested synchronic progression away from prenasalized stops.

Zamma (pp. 157-176) examines the connection between accentuation and rendaku in Japanese surnames, referring, like Kubozono, to Sugito (1965), who shows that words that undergo rendaku tend to be accentless. This trend is apparent, for example in surnames ending in *ta* ('field'). Zamma (like Kubozono, this volume) also observes the influence of the onset consonant of the last mora of the preceding morpheme on rendaku. Takayama (pp. 177-190) discusses loanwords that are perceived by Japanese speakers as native words and undergo rendaku voicing. Some such words phonotactically resemble native words. Takayama also discusses Sino-Japanese words that do not phonotactically resemble native words but pattern

phonologically with native words. He thus divides the lexicon by phonological rather than strictly etymological criteria.

Suzuki's article (pp. 191-204) reports on research in applied computational phonology in the area of speech recognition. Specifically, Suzuki seeks to improve the performance of a speech recognition engine with respect to numeral-classifier combinations that are subject to morphologically conditioned allomorphy, as well as context-independent free variation. His speech recognition engine operates stochastically, by seeking to find an output sentence that has the highest probability of matching an acoustic input string.

Maekawa and Kikuchi (pp. 205-228) argue, from experimental evidence, that devoicing of Japanese vowels is probabilistic rather than completely predictable. They analyze a large corpus of read speech that was annotated with respect to the context in which devoicing occurs. They find that all vowels, not just /i/ and /u/, experience some degree of devoicing in certain contexts. Devoicing is affected by (a) speaking rate, (b) vowel length, (c) avoidance of consecutive devoicing (e.g. *CVCVC...) and (d) sonorancy, manner of articulation, and gemination of adjacent consonants.

Kondo (pp. 229-245) examines the acoustic effects of syllable structure on vowel devoicing. She compares how vowel devoicing affects: (a) durations of morae, (b) stop/affricate closures and (c) vowel intensities. She also argues that devoiced vowels cannot form a syllable nucleus, thus causing adjacent consonants to resyllabify and a word's mora count to be reduced. This hypothesis accounts for the avoidance of consecutive devoiced vowels, which would render consonants unsyllabifiable. Kondo does not discuss independent evidence for mora reduction. If devoicing results in demorafication, we would predict that mora-count-sensitive processes would be affected by vowel devoicing – for example the effects of moraic length on compound accentuation (Poser 1990, Kubozono 1995).

Sugito (pp. 247-260) examines devoicing of high vowels in Osaka Japanese, where, unlike in the Tokyo dialect, devoiced vowels are often accented. In an earlier study, Sugito (1969) had found that falling contours on vowels following a devoiced vowel could cause the preceding devoiced vowel to be perceived as accented. Sugito conducted both acoustic and physiological experiments on Osaka speakers pronouncing words with potentially devoiceable vowels at different speeds. She found that accented, devoiced vowels were more likely to occur in faster speech. She also found, through measurements of crycothyroid and sternohyoid activity, that devoiced vowels were not only perceived as if accented, but were also produced with laryngeal activity associated with pitch accent.

Tanaka's paper (pp. 261-276) proposes a fixed hierarchy of phonological prominence:

(1) accent > tone > sonority > voicing.

Each element in the hierarchy is composed of a set of phonetic factors that is a proper subset of those elements that dominate it: for example, accent is composed of

{vibration, aperture, pitch/duration, intensity}, tone/length only of the first three, sonority only of the first two, and voicing only of the first, i.e., vibration. This hierarchy captures implicational relations (for example that sonorancy implies voicing but not vice versa). Tanaka applies this model to interactions between pairs of elements in (1): for example, the interaction between accent and voicing. He then looks at how voicing and accent in Japanese conspire to maximize syllable prominence, in cases where high vowel devoicing is blocked so that accent and voicing can coincide, or where accent shifts away from a default position to avoid coincidence with a devoiced vowel. Finally, he looks at cases of optional devoicing and accent shift. He accounts for optionality in the OT framework through free ranking of constraints (Antilla 1997, 2002), where a non-shifted form is optimal through "sympathy" (McCarthy 1999, 2003) to a candidate that does not devoice or shift.

In summary, although the articles cover a variety of topics, certain issues recur throughout the book – for example, the phonological vs. phonetic nature of voicing, the stratifcation of the Japanese lexicon, the interaction between voicing, nasality and OCP effects, and the interaction between pitch accent and voicing of both vowels and consonants. Overall, the articles are well-argued and presented, and offer an interesting complexity of data and issues. They also offer a sense that there continues to be a significant amount of research to be done in the phonology and phonetics of voicing in Japanese. In this light, „Voicing in Japanese" constitutes an important contribution to Japanese phonology and phonetics.

References

Antilla, A. 1997. Deriving variation from grammar. In Hinskens, F., van Hout, R. and Wetzels, W. L. (eds.): *Variation, change, and phonological theory*. Amsterdam: John Benjamins. 35-68.

Antilla, A. 2002. Morphologically conditioned phonological alternations. *Natural Language and Linguistic Theory*, 20: 1-42.

Itô, J., Mester, R.-A. 1986. The phonology of voicing in Japanese. *Linguistic Inquiry*, 17: 49-73.

Itô, J., Mester, R.-A. 1995. Japanese phonology. In Goldsmith, J. (ed.): *The handbook of phonological theory*. Blackwell. 817-838.

Itô, J., Mester, R.-A. 1995. The core-periphery structure of the lexicon and constraints on reranking. *University of Massachusetts Occasional Papers in Linguistics* 180.

Itô, J., Mester, R.-A. 1998. *Markedness and word structure: OCP effects in Japanese*. Ms. University of California, Santa Cruz.

Itô, J., Mester, R.-A. 1999. The structure of the phonological lexicon. In Tsujimura, N. (ed.): *The handbook of Japanese linguistics*. Malden, MA and Oxford, U.K.: Blackwell Publishers. 62-100.

Itô, J., Mester, R.-A. and Padgett, J. 1995 Licensing and redundancy: underspecification in Optimality Theory. *Linguistic Inquiry*, 26: 571-614.

Itô, J., Mester, R.-A. and Padgett, J. 1999. Lexical classes in Japanese: A reply to Rice. *Phonology at Santa Cruz*, 6: 39-46.

Kaye, J. D., Lowenstamm, J. and Vergnaud, J.-R. 1990 Constituent structure and government in phonology. *Phonology*, 7: 193-231.

Kubozono, H. 1995. Constraint interaction in Japanese phonology: Evidence from compound accent. *Phonology at Santa Cruz*, 4: 21-38.

Kubozono, H. 1999. Mora and syllable. In Tsujimura, N. (ed.): *The handbook of Japanese linguistics*. Blackwell. 31-61.

McCarthy, J. J. 1999. Sympathy and phonological opacity. *Phonology*, 16: 331-399.

McCarthy, J. J. 2002. Sympathy, cumulativity, and the Duke-of-York gambit. In Féry, C. and van de Vijver, R. (eds.): *The optional syllable*. Cambridge: Cambridge University Press.
McCawley, J. D. 1968. *The phonological component of a grammar of Japanese.* The Hague: Mouton.
Mester, R.-A. and Itô, J. 1989. Feature predictability and underspecification: Palatal prosody in Japanese mimetics. *Language,* 65: 258-93.
Okumura, M. 1955. Rendaku. In Kokugo Gakkai (ed.): *Kokugogaku jiten*. Tokyo: Tôkyôdô. 916-961.
Otsu, Y. 1980. Some aspects of rendaku in Japanese and related problems. In Otsu, Y. and Farmer, A. (eds.): *Theoretical issues in Japanese linguistics: MIT Working Papers in Linguistics 2*. 207-227.
Pater, J. V. 1999. Austronesian nasal substitution and other NC effects. In Kager, R. W. J., van der Hulst, H. G. and Zonneveld, W. (eds.): *The prosody-morphology interface*. Cambridge: Cambridge University Press. 310-343.
Poser, W. J. 1990. Evidence for foot structure in Japanese. *Language,* 66: 78-105.
Prince, A. S. and Smolensky, P. 1993. *Optimality Theory: Constraint interaction in generative grammar*. Ms. Rutgers University and University of Colorado. [To appear, Blackwell, Oxford]
Sugito, M. 1965. Shibata-san to Imada-san: Tango-no chookakuteki benbetsu ni tsuite-no ichi koosatsu [Mr. Shiba-ta and Mr. Ima-da: A study in the auditory differentiation of words], *Gengo Seikatsu* 165 [S40-6]: 64-72 [Reproduced in Miyoko Sugito (1998) *Nihongo Onsei no Kenkyu* (Studies on Japanese Sounds), Izumi-Shoin, Vol. 6, 3-15.]
Sugito, M. 1969. Akusento no aru museika boin [A study on accented voiceless vowels]. *The Bulletin,* 132 (The Phonetic Society of Japan): 1-3.
Tateishi, K. 2001. On'in jisho kurasu seeyaku no bunpu ni tsuite [On the distribution of constraints for phonological sub-lexica]. Paper presented at the 26th Annual Meeting of the Kansai Linguistic Society, Ryukoku University, Kyoto.
Vance, T. 1980. Comments on some aspects of rendaku in Japanese and related problems. In Otsu, Y. and Farmer, A. (eds.): *Theoretical issues in Japanese linguistics. MIT Working Papers in Linguistics 2*. 229-236.
Vance, T. 1996. Sequential voicing in Sino-Japanese. *Journal of the Association of Teachers of Japanese,* 30: 22-43.

Gábor Olaszy: Mássalhangzó-kapcsolódások a magyar beszédben [Consonant Clusters in Hungarian Speech]

Budapest: Tinta Könyvkiadó, 2007 (265 pp. ISBN 978-963-7094-77-4)

Reviewed by: **Péter Siptár**
Eötvös Loránd University, Budapest, Hungary
e-mail: siptar@nytud.hu

This book is groundbreaking in its genre as it presents the first comprehensive analysis of the acoustic structure of Hungarian (intervocalic) consonant clusters ever published in Hungarian (or, for that matter, in any language). The book consists of a preface, a short introduction, six core chapters, a brief list of references, and an appendix.

The first chapter (pp. 13–21) introduces the material and research method. Four cluster types are discussed in the book from CC to CCCCC, involving a total of 897 different combinations of consonants. First, a word list containing each consonant cluster in this language was provided, as well as an example of each cluster being used in a word (e.g., *la**bd**a* 'ball'; *elle**ntm**ondó* 'contradictory'; *sü**ltkr**umpli* 'fried potatoes'; *nye**lvstr**uktúra* 'linguistic structure'). The word list was recorded by five

male and five female speakers as they read. This recording then served as the target of a meticulous acoustic analysis. The sound files were semi-automatically labeled; i.e., voiced/voiceless markers and sound boundaries were added to the speech waveform and a phonemic transcription was also provided for each item. As it was impossible to include acoustic diagrams of all 897 words in a relatively slender book, a webpage was developed where the acoustic diagrams of the whole word list are stored as uttered by one male and one female speaker (http://fonetika.nytud.hu/cccc). This speech database, as the author points out, is an integral part of the book. The text of the book and the webpage material, taken together, provide the reader with complex and comprehensive information about Hungarian intervocalic consonant clusters. In addition, the book contains numerous figures, illustrating the structural changes of sounds observable when produced clusters. Computer-based support was used both for the analysis and for making representative figures about the acoustic changes that the various consonants undergo when they occur in clusters.

The second chapter (pp. 23–143) discusses the basic cluster types, in which two consonants make up sound combinations. All CC clusters (373 types) are discussed. Duration characteristics, spectral features, and structural changes are exaplained. Here are some examples of the results: in CC clusters, the duration of both Cs is shorter than the duration of the same consonants in intervocalic position. The complexity of articulatory movements is also reflected by duration (less complex movements result in shorter duration and vice versa). In the most complex cases, additional sound elements appear in the waveform around the meeting point of the two consonants (spirantization, schwa element, coarticulatory silent phase, etc.).

The third chapter (pp. 145–193) discusses CCC clusters; 445 types of such sound combinations are investigated. The author compares these clusters with the relevant CC clusters, and describes the differences (if any). The shortening tendency continues, i.e., the consonants have (even) shorter durations here than in CC clusters. An interesting result is the frequency distribution of consonants occurring as building elements of the CCC clusters. The most frequent sound is the tremulant, and it occurs only in initial and final position.

In chapter four (pp. 195–208), the author turns to CCCC clusters; 74 types are discussed. The shortening tendency ceases to apply in these clusters: the sound durations are similar to those in CCC clusters. To explain this phenomenon, the author hypothesizes that the articulation of four consonants in a row is so complicated that one cannot do it as easily and automatically as in the CC and CCC cases. The most frequent building element is again the tremulant: it occurs in the initial and final positions as seen earlier.

Chapter five is a short one (pp.209–211), containing the description of CCCCC clusters. These clusters are very rare in speech. Altogether, five types were found.

In chapter six (pp. 213–217), the author introduces the accompanying webpage (mentioned above) and the interactive program that allows searching for any cluster in the speech database. Two options are available: search on the sound level and on the text level. The following acoustic features can be displayed: waveform with

sound boundaries, spectrograms, and intensity patterns. It is also possible to listen to the examples. The user can measure formant frequency and intensity values and compare the acoustic structures of pairs of words.

The appendix (pp. 221–265) contains lists of frequency distributions for all cluster types, lists with detailed duration values, as well as alphabetical lists, according to the position of the consonant inside the cluster.

The book is intended primarily for use as a textbook in university courses, but secondaryschool teachers of Hungarian grammar can also employ it in their work; and indeed it can be recommended to anyone interested in the structure of human speech (and Hungarian speech in particular). The rich material presented here can be used as valuable input for linguists, speech researchers, communication engineers, and experts in speech technology in their research activities.

Hangidőtartamok és időszerkezeti elemek a magyar beszédben [Sound durations and temporal factors in Hungarian speech]

Budapest: Akadémiai Kiadó, 2006 (188 pp., ISBN 978 963 05 8437 1)

Reviewed by: **Kálmán Abari**
Department of Psychology, University of Debrecen, Hungary
e-mail: abari.kalman@gmail.com

This book describes research supported by computer and speech databases which examines Hungarian sound durations and other temporal factors in speech. It was published in the series of *Nyelvtudományi Értekezések* [Papers in Linguistics, the 155th issue] by Akadémiai Kiadó, in Budapest.

The book contains four chapters and an appendix (for a total of 192 pages). In the appendix (60 pages), the author provides tables of the specific sound durations for Hungarian as a function of the preceding and following sound, as well as a description of the measurement tool. This software was developed by Kálmán Abari (a mathematician at the Debrecen University, Hungary).

The first chapter revealed new methods for the measurement of sound durations by computer. The Hungarian sound types were discussed, the question of sound and word boundary was described, and two methods were presented for the measurement of these boundaries. The first method is new and based on perceptual judgements, i.e. whether the sound is too long or too short in an utterance. For this method, he used specially generated synthetic speech. The second method was the traditional one, i.e., direct measurements on the wave form, but supported by the described measuring tool. This measurement used a speech database with pre-deteremined sound boundary markers (i.e., hand made).

The second chapter described a model for the prediction of the duration of speech sounds from text. This is the first text-to-duration model designed for Hungarian.

The model had a down-up structure, and the final durations were created along three levels. The basic ground was represented by the specific durations, the second (intermediate level) was represented by the word rhythmn, and the highest level was the suprasegmental level. The model was based on rules snce all types of Hungarian utterances can be handled on sentence level (short, long, simple, complex, statements, questions etc.). The new method for the definition of sound duration was described in detail in this chapter. The results of the model were evaluated with direct comparison of natural and synthesised sentences. The accuracy was 95%. The preferred Hungarian speech synthesizer (Profivox) uses this model for speech synthesis.

The third chapter contained data from direct duration and time structure measurements. Sound and word duration data were described and the significant effects of surrounding sounds were also investigated. A comparison among readings of different text types like news reading, prose performance, tales, advertisements and narrator speech (in films about the nature) was completed.

Using these comparisons, the author revealed that there were differences among the reading styles and described this difference algorithmically. Based on this research the Profivox speech synthesizer was equipped with this stylistic knowledge so that it can produce different speaking styles, depending on the type of text.

In the fourth chapter, the author compared the main time structure components in read and spontaneous speech. He indicated that there was no significant difference between segmental sound durations in read and spontaneous speech (i.e., a dialogue taken from a radio broadcast). The measurable difference between the two speaking styles resided in the break strategy (the location and length of breaks used by the speaker).

In the Appendix, the specific sound duration tables were displayed and a description of the measurement program was described. An internet presentation of the results can be found at url: http://fonetika.nytud.hu/hitint. A database for the presentation of sound duration-maps of Hungarian words is presented there (1.5 million word forms).

In general, the book's goal was to present current time structure data for Hungarian speech sounds. The results highlighted new data while verifying former acoustic measurements. The results would be useful for linguists, phoneticians and for those who are learning/working in the domain of speech technology in general.

Jacob Benesty, M. M. Sondhi and Yiteng Huang (eds.):
Handbook of speech processing

Springer, 2008 (xxxvi, 1176 p., 456 illus. in color, with DVD, Hardcover,
ISBN: 978-3-540-49125-5)

Reviewed by: **Eduardo López Gonsalo**
Technical Universty of Technology , Madrid, Spain
e-mail: eduardo.lopez@upm.es

This book is a comprehensive overview of most of the major topics associated with speech processing written by the most renowned authors in each topic. The book is well structured with a clearly organized topics. It is intended for use by the researcher more than the student. The book is organized in nine sections that cover all current speech applications.

The first two sections cover the fundamental theories that should be understood before conducting an in-depth study of speech processing. Separating background theory from its use is also useful in that it allows a rigorous approach to its description. The inclusion of recommended further readings, in addition to the vast number of references appearing in each chapter, make the book a very good starting point for anyone doing work in speech processing.

The third section concerns itself with speech coding topics. In this Section, the reader will find everything from FFTs to multi-rate signal processing and speech signal representations to speech coding. Again, this section is well written and the reader is not forced to refer to other texts to understand what is written. If a topic is not expanded upon here, then it is an indication that is not dealt further with in any great depth in the remainder of the book.

The fourth section of the book covers text-to-speech synthesis form, including the principles and rule or corpus generation techniques, as well as a completed linguistic analysis of these systems. In addition, new techniques, such as voice conversion and emotion generation are given. It shows numerous block diagrams of the items that you need in order to build such a system. For instance, it provides numerous algorithms in pseudocode. It also dedicates a subsection to each block of the text-to-speech system block diagram.

The fifth section is about speech recognition. This section is very thorough in its treatment of the subject. It starts with historic background information and immediately follows with a discussion of Hidden Markov Models, which is almost exclusively the method employed in the pattern matching stage of speech recognition. Any algorithms mentioned are thoroughly discussed, which really makes the book useful. In fact, algorithms are presented throughout the book, making it a practical reference, as much as a theoretical one. This is important because there is a big jump from understanding theory to being able to implement an algorithm to demonstrate that theory. Another topic covered includes an excellent

chapter on environmental robustness. Language modelling and search algorithms also are discussed in detail.

Section six, on speaker recognition, is a good overview of the field and better than any book I've seen on the subject. It discusses in detail what you would need to do in order to implement that particular approach.

Section seven on is on language recognition, and it follows a similar structure as section six, given that the implementation usually overlaps with the previous topic.

The eighth section is about speech enhancement. which is rarely treated in a book about speech processing as a unique chapter. Here, as in the rest of the book, an encyclopaedic approach is provided, with topics on every major application.

In conclusion, I would highly recommend that anyone interested in speech processing have a copy of this encyclopaedic work.

MEETINGS, CONFERENCES AND WORKSHOPS

2008

11–12 January 2008
Budapest Uralic Workshop 6 (BUW 6)
Budapest (Hungary)
http://www.nytud.hu/bum6 (from 20 Aug 2007)
sipos@nytud.hu

11–14 January, 2008
The 6th Annual Hawaii International Conference on Arts and Humanities
Honolulu, Hawaii (USA)
http://www.hichumanities.org/
humanities@hichumanities.org

16–17 January, 2008
1st National Seminar on Speech and Language Disorders: Assessment and Intervention in the Indian context
Berhampur, India
gouriraj@sancharnet.in

17–18 January, 2008
CUNY Phonology Forum Conference on the Syllable
New York (USA)
syllable@cunyphonologyforum.net

18 January, 2008
The 2nd Czecho-Slovak Conference of the International Society of Phonetic Sciences
Prague (Czech Republic)
dubeda@ff.cuni.cz

23 January, 2008
Corpora in Phonological Research
Toulouse (France)
anne.przewozny@univ-tlse2.fr

24–26 January, 2008
Old World Conference in Phonology 5
Toulouse (France)
anne.przewozny@univ-tlse2.fr

9–10 February, 2008
The 1st Nordic Conference of Clinical Linguistics (NorConfClinLing2008)
Joensuu (Finland)
http://cc.joensuu.fi/linguistics/NorConfClinLing2008/
NorClinLing2008@joensuu.fi

18–21 February 2008
Le Changement Linguistique et ses Théories (ED-M3-2008)
Fribourg (Switzerland)
http://www2.unine.ch/structuration_periodes
mathieu.avanzi@unine.ch

22–23 February, 2008
Current Approaches to Spanish & Portuguese Second Language Phonology
Minneapolis (USA)
http://spanport.cla.umn.edu/L2phonology
facex002@umn.edu

22–24 February, 2008
Penn Linguistics Colloquium (PLC 32)
Philadelphia (USA)
http://www.ling.upenn.edu/Events/PLC/
plc32@ling.upenn.edu

27 February, 2008
Methodological Aspects of Intonation Research
Bamberg (Germany)
pia.bergmann@germanistik.uni-freiburg.de

27–29 February, 2008
A Comparison of Signed and Spoken Languages
Bamberg (Germany)
wrobel@daf.uni-muenchen.de

28–29 February, 2008
LangTech 2008 Conference
Rome (Italy)
http://www.langtech.it/en/
secretariat@langtech.it

28–29 February, 2008
The Role of Phonology in Reading Acquisition
Bamberg (Germany)
kathrin_schrader@gmx.de

6–8 March, 2008
English as a Lingua Franca (ELF Forum)
Helsinki (Finland)
http://www.eng.helsinki.fi/ELFforum/
ELF-Forum@helsinki.fi

7–8 March, 2008
African American Women's Language Conference '08
San Antonio, Texas (USA)
sonja.lanehart@utsa.edu

25 March, 2008
Categorical Phonology and Gradient Facts (GLOW Phonology Workshop)
Newcastle upon Tyne (UK)
glow31@ncl.ac.uk

27–28 March, 2008
Belgrade International Meeting of English Phoneticians (BIMEP08)
Belgrade (Serbia)
biljana.cubrovic@gmail.com

27–28 March, 2008
The 6th International Conference on Informatics and Systems – Special Track On Natural Language Processing (INFOS 2008)
Cairo (Egypt)
http://www.fci.cu.edu.eg/INFOS2008/

27–28 March, 2008
3rd Northern Englishes Workshop (NEW 3)
Salford (UK)
http://www.esri.salford.ac.uk/esri/m/?s=4
P.Tipton@salford.ac.uk

28 March, 2008
Workshop on Empirical Approaches to Speech Rhythm
London (UK)
http://www.phon.ucl.ac.uk/rhythm2008
rhythm2008@phon.ucl.ac.uk

28 March, 2008
The Neurocognition of Memory and Language
Washington DC (USA)
http://cbbc.georgetown.edu/workshops/2008RA.html
cbbc@georgetown.edu

2–4 April, 2008
14. Arbeitstagung zur Gesprächsforschung
Mannheim (Germany)
http://www.gespraechsforschung.de/tagung/programm.htm
tagung@gespraechsforschung.de

2–5 April, 2008
Didactique du français par la pratique théâtrale
Tunis (Tunisia)
colloquetunis@googlemail.com

3–5 April, 2008
5. Interdisziplinäre Tagung über Sprachentwicklungsstörungen (ISES 5)
Mainz (Germany)
ses5@fh-fresenius.de
http://ises5.fh-fresenius.de

3–5 April, 2008
Sociolinguistic Issues in the Use of Language
Amsterdam (The Netherlands)
http://www.taalstudio.nl

4–5 April, 2008
Phonology, Syntax and the Lexicon: Interdependence (ALOES 2008)
Paris (France)
http://www.univ-paris13.fr/CRIDAF/avril2008.htm
nballier@free.fr

5 April, 2008
SFU Phonology Fest 2008
Burnaby (Canada)
http://www.sfu.ca/linguistics/phonfest/phonfest_index.html
kns3@sfu.ca

11–13 April, 2008
Experimental and Theoretical Advances in Prosody
Ithaca (USA)
http://ling.cornell.edu/prosody08
prosody08@gmail.com

18–20 April, 2008
The 8th Phonetic Conference of China (PCC2008)
Beijing (China)
http://www.pcc2008.cn/PCC2008_en/index.html
phonetics2008@gmail.com

25–26 April, 2008
The 3rd "Talking Across the World" Conference 2008 (TAW 3)
Bangalore (India)
http://www.talkingacrosstheworld.org
egconf@polyu.edu.hk

2–3 May, 2008
Linguistic Variation Across the Lifespan
Columbus (USA)
springsym@ling.osu.edu
http://www.ling.ohio-state.edu/~springsym/

5 May, 2008
First EMUS Conference – Expressivity in Music and Speech
Campinas (Brazil)
http://recherche.ircam.fr/equipes/analyse-synthese/EMUS

6–9 May, 2008
Speech Prosody 2008
Campinas (Brazil)
http://www.sp2008.org/
sp2008_info@iel.unicamp.br

5–7 May, 2008
The International Workshop on Spoken Languages Technologies for Under-Resourced Languages (SLTU)
Hanoi (Vietnam)
http://www.mica.edu.vn/sltu

9–11 May, 2008
Phonetics of the Slavic Languages
New Haven (USA)
odi.reich@yale.edu

9–11 May, 2008
5th North American Phonology Conference (NAPhC5)
Montreal (Canada)
http://linguistics.concordia.ca/naphc5/
cogsci@alcor.concordia.ca

17–18 May, 2008
The 2nd Conference on Language, Discourse & Cognition (CLDC-2)
Taipei (Taiwan)
cldc2008@ntu.edu.tw; d94142001@ntu.edu.tw
http://homepage.ntu.edu.tw/~gilntu/

22–24 May, 2008
16th Manchester Phonology Meeting (16MFM)
Manchester (UK)
http://www.englang.ed.ac.uk/mfm/16mfm.html
patrick.honeybone@ed.ac.uk

28–30 May, 2008
Language Resources and Evaluation Conference
Marrakech (Morocco)
http://www.lrec-conf.org/lrec2008/

31 May – 1 June, 2008
2008 Conference of Japan Second Language Acquisition (J-SLA2008)
Kyoto (Japan)
http://www.j-sla.org/e/index.html
shunjil2@yahoo.co.jp

4–5 June, 2008
4th Intercultural Rhetoric and Discourse Conference (IR Conference)
Indianapolis (USA)
http://www.iupui.edu/~icic/IRconference.htm
uconnor@iupui.edu

9–13 June, 2008
XXVIIes Journées d'Étude sur la Parole (JEP'08)
15e Conférence sur le Traitement Automatique des Langues Naturelles (TALN'08)
10e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'08)
Avignon (France)
http://www.lia.univ-avignon.fr/jep-taln08/
jean-francois.bonastre@univ-avignon.fr

12–13 June, 2008
The 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2008)
Barcelona (Spain)
http://www.iceis.org/workshops/nlpcs/nlpcs2008-cfp.html

13–14 June, 2008
Prosody-Syntax Interface Workshop (PSI 2)
Berlin (Germany)
http://www.zas.gwz-berlin.de/events/psi2/
downing@zas.gwz-berlin.de

16–18 June, 2008
Perception and Interactive Technologies for Speech-Based Systems (PIT08)
Kloster Irsee (Germany)
http://it.e-technik.uni-ulm.de/World/Research.DS/irsee-workshops/pit08/introduction.html

17–18 June, 2008
Prosody of Expressivity in Speech and Music
Paris (France)
http://recherche.ircam.fr/equipes/analyse-synthese/EMUS

19–20 June, 2008
ACL Special Interest Group in Morphology and Phonology (SIGMORPHON 2008)
Columbus, Ohio (USA)
http://phonology.cogsci.udel.edu/sigmorphon2008/
sigmorphon2008@udel.edu

25–28 June, 2008
12th Congress of the International Clinical Phonetics and Linguistics Association
Istanbul (Turkey)
http://www.icpla2008.org
dilkom@anadolu.edu.tr

26–28 June, 2008
Structural Features of Varieties of French in Contact (PCF)
New Orleans (USA)
klingler@tulane.edu

26–28 June, 2008
First SignTyp Conference
Storrs, CT (USA)
http://linguistics.uconn.edu/sign/
signtyp@uconn.edu

29 June – 4 July, 2008
Acoustics'08 Paris
Paris (France)
http://www.acoustics08-paris.org/

30 June – 4 July, 2008
Introduction à la Phonologie Déclarative
Porto (Portugal)
jveloso@letras.up.pt

30 June – 2 July, 2008
11th Laboratory Phonology Conference
Wellington (New Zealand)
http://www.vuw.ac.nz/labphon11
labphon11@vuw.ac.nz

7–8 July, 2008
The Nature and Development of L2 French
Southampton (UK)
http://www.flloc.soton.ac.uk/conferences.html
annabelle.david@ncl.ac.uk

9 July, 2008
The Phonological Deficit Hypothesis
Asheville (USA)
http://www.triplesr.org/conference
braze@haskins.yale.edu

9–11 July, 2008
Rencontre « André Martinet, linguiste »
Paris (France)
f.i.dhiver@wanadoo.fr

9–11 July, 2008
Computers Helping People with Special Needs
Linz (Austria)
http://cwst.icchp.org/node/13

9–12 July, 2008
Congrès mondial de linguistique française (CMLF-08)
Paris (France)
http://www.ilf.cnrs.fr/spip.php?rubrique4
benoit.habert@ens-lsh.fr

10–12 July, 2008.
Choice for Voice 2008
London (UK)
www.british-voice-association.com

12–15 July, 2008
International Conference on Machine Learning and Cybernetics (ICMLC 2008)
Kunming (China)
http://www.icmlc.com/welcome.htm

13–16 July, 2008
International Professional Communication Conference (IPCC)
Montréal (Canada)
IPCC2008@gmail.com

21–26 July, 2008

Workshop on Speech Sciences in Linguistics (CIL 18)
Seoul (South Korea)
http://www.cil18.org/workshop/workshop_04.htm
conference@speechsciences.org

28 July – 1 August, 2008

The 11th International Congress for the Study of Child Language (IASCL 2008)
Edinburgh (UK)
http://www.in-conference.org.uk/IASCL/
IASCL@in-conference.org.uk

31 July – 2 August, 2008

Consonant Clusters and Structural Complexity
Munich (Germany)
http://www.phonetik.uni-muenchen.de/cluster
cluster@phonetik.uni-muenchen.de

4–6 August, 2008

10th Nordic Prosody conference
Helsinki (Finland)
www.helsinki.fi/speechsciences/np2008/

4–7 August, 2008

Language, Communication and Cognition
Brighton (UK)
http://www.languageandcognition.net
LCC@Brighton.ac.uk

4–8 August, 2008

13th International Conference on Methods in Dialectology (Methods XIII)
Leeds (UK)
http://www.leeds.ac.uk/english/methods.htm
engmeth@leeds.ac.uk

10–16 August, 2008

VISPP Summer School 2008
Kuressaare (Estonia)
http://vispp2008@phon.ioc.ee
evispp2008@phon.ioc.ee

18–22 August 2008

22nd International Conference on Computational Linguistics
Manchester (UK)
http://www.coling2008.org.uk/

18–22 August, 2008

Summer School on Corpus Phonology
Augsburg (Germany)
http://www.uni-augsburg.de/summerschool
summerschool08@phil.uni-augsburg.de

25–27 August, 2008
Special Session on Greek Phonetics (ExLing2008)
Athens (Greece)
http://www.exling.gr/2008/home2008.htm
exling@phil.uoa.gr

1–4 September, 2008
Intensive Workshop: Intonation in English
Kowloon (Hong Kong)
http://www.engl.polyu.edu.hk/events/intonworkshop2008/
egclaw@inet.polyu.edu.hk

3–5 September, 2008
AFLS Conference
Oxford (UK)
http://www.afls.net/conferences.html
afls2008oxford@hotmail.fr

8–12 September, 2008
11th International Conference on Text, Speech and Dialogue (TSD 2008)
Brno (Czech Republic)
http://www.tsdconference.org
tsd2008@tsdconference.org

10–13 September, 2008
Eurosla 18
Aix-en-Provence (France)
http://blog. univ-provence. fr/blog/eurosla18

11–12 September, 2008
Workshop on Phonological Variation in Voicing
Amsterdam, Leiden (The Netherlands)
Marc.van.Oostendorp@Meertens.KNAW.nl

15–17 September, 2008
Third TIE Conference on Tone and Intonation (TIE3)
Lisbon (Portugal)
http://www.fl.ul.pt/LaboratorioFonetica/TIE3/
sonia.frota@mail.telepac.pt

18–20 September, 2008
Discourse Coherence – Text and Theory
Paris (France)
http://www.celta.paris-sorbonne.fr/
celta@paris-sorbonne.fr

20 September, 2008
Comparing Prosodies Grammatically
Cambridge, MA (USA)
http://www.fas.harvard.edu/~lingdept/comparingProsodiesgrammatically.html

22–26 September, 2008
Interspeech (ICSLP) 2008
Brisbane (Australia)
http://www.isca-speech.org/call4prop_interspicslp2008.htm

26–28 September, 2008
Laboratory Approaches to Spanish Phonology (SP4)
Austin (USA)
http://www.utexas.edu/cola/conferences/lasp/main/
moll@mail.utexas.edu

26–28 September, 2008
Living, Working and Studying in Vehicular Languages
Turku (Finland)
http://www.hum.utu.fi/oppiaineet/ranskankieli/tutkimus/konferenssit/vehicular.html
freder@utu.fi

28 September, 2008
Psycholinguistics in Teaching English as a Second Language (TESOL)
Reading (UK)
j.c.field@reading.ac.uk

2–5 October, 2008
7. Internationale Stuttgarter Stimmtage
Stuttgart (Germany)
www.gesprochenes-wort.de

6–8 October, 2008
Acoustics Week in Canada
Vancouver (Canada)
http://www.caa-aca.ca
mhodgson@interchange.ubc.ca

15 October, 2008
Les Universaux prosodiques
Paris (France)
http://www.umr7023.cnrs.fr/
mrusso@univ-paris8.fr

15–18 October, 2008
2nd International Conference on Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions
and
18th Czech-German Workshop on Speech Processing
Prague (Czech Republic)
http://www.ufe.cz/events/cost2102.php

23–27 October, 2008
Instrumental Phonology: Patterns and Variation
México City (Mexico)
http://lef.colmex.mx
eherrera@colmex.mx

31 October – 2 November, 2008
Boston University Conference on Language Development (BUCLD)
Boston, MA (USA)
http://www.bu.edu/linguistics/APPLIED/BUCLD/
langconf@bu.edu

4–6 November, 2008
Applications of Phonetics and Phonology On Arabic
Mafraq (Jordan)
http://www.aabu.edu.jo/art/home.htm
said@aabu.edu.jo, said19681@yahoo.com

7–9 November, 2008
Experimental Methods in Language Acquisition Research (EMLAR)
Utrecht (The Netherlands)
http://www.let.uu.nl/emlar/
emlar@let.uu.nl

13–16 November, 2008
50th Annual M/MLA Convention
Minneapolis (USA)
http://www.uiowa.edu/~mmla
smburt@ilstu.edu

24–26 November, 2008
Congress of Phonetics and Phonology
Niteroi, Rio de Janeiro (Brazil)
http://sbfonetica.vilabol.uol.com.br
mtmatta@terra.com.br

27–29 November, 2008
Prosodic Interface Relations (PIR 2008)
Stuttgart (Germany)
http://www.ims.uni-stuttgart.de/veranstaltungen/pir2008
pir2008-info@ims.uni-stuttgart.de

3–5 December, 2008
International Symposium: 30 Aniversari del Laboratori
Barcelona (Spain)
http://ub.edu/labfon/simposiumc.htm
dszmidt@ub.edu

4 December, 2008
Heard around the World
Brussels (Belgium)
http://homepages.ulb.ac.be/~heard
heard@ulb.ac.be

4–5 December, 2008
1st International Workshop on Cataloguing and Coding of Spoken Language Data (CatCod 2008)
Orléans (France)
http://www.catcod.org

7–12 December 2008
E-Humanities – an Emerging Discipline
Indianapolis (USA)
http://www.clarin.eu/
clarin@clarin.eu

8–12 December, 2008
International Seminar on Speech Production (ISSP'2008)
Strasbourg (France)
http://issp2008.loria.fr/

15–16 December, 2008
2nd International Symposium on Universal Communication
Osaka (Japan)
http://www.is-uc.org/2008/

2009

5–7 January, 2009
Experimental Studies on Intonation: Phonetic, Phonological and Psycholinguistic Aspects of Sentence
Potsdam (Germany)
http://www.ling.uni-potsdam.de/pip/daten/workshop.html
gerrit@ling.uni-potsdam.de

15–17 January, 2009
Conference on the Foot in Phonology
New York (USA)
http://www.cunyphonologyforum.net/foot.php
foot@cunyphonologyforum.net

16–17 January, 2009
The Divison of Labour between Morphology and Phonology
Amsterdam (The Netherlands)
http://www.uni-leipzig.de/~exponet/meet.htm
doreengeorgi@gmx.de

21–24 January, 2009
6th Old World Conference in Phonology (OCP6)
Edinburgh (UK)
http://www.lel.ed.ac.uk/ocp6
patrick.honeybone@ed.ac.uk

4 March, 2009
Insertions and Deletions in Speech
Osnabrück (Germany)
insertions@zas.gwz-berlin.de

4–6 March, 2009
DGfS Workshop "Rhythm beyond the word" (DGfS-AG RBW)
Osnabrück (Germany)
ruben@ling.uni-potsdam.de

26–27 March, 2009
Workshop on Pharyngeals & Pharyngealisation
Newcastle upon Tyne (United Kingdom)
http://www.ncl.ac.uk/linguistics/news/events/item/international-workshop-on-pharyngeals-pharyngealisation
ghada.khattab@ncl.ac.uk

27 March, 2009
Regards croisés sur la prosodie du français
Paris (France)
http://www2.unine.ch/conscilaprosodie/page26423.html
mathieu.avanzi@unine.ch

9–10 April, 2009
International Conference on Prosody and Iconicity (ProsIco 2009)
Rouen (France)
http://www.prosico2009.com
hancilfr@yahoo.fr

23–25 April, 2009
Experimental Pragmatics Conference 2009 (XPrag 2009)
Lyon (France)
http://xprag.l2c2.isc.cnrs.fr/XPrag/
cchevallier@isc.cnrs.fr

1–3 May, 2009
3rd Brazilian Bilingual Conference
São Paulo (Brazil)
http://www.playpen.com.br/congresso1bilingue.asp?version=english
playpen@uol.com.br

28–30 May, 2009
21st International Conference on Foreign and Second Language Acquisition (ICFSLA 2009)
Szczyrk (Poland)
http://uranos.cto.us.edu.pl/~icfsla/contact.htm
szczyrkconference@op.pl

3–5 June, 2009
English Pronunciation: Issues & Practices (EPIP)
Chambéry (France)
colloque-epip@univ-savoie.fr

5 June, 2009
Nasal 2009
Montpellier (France)
http://w3.umh.ac.be/~nasal/Workshop/Englishversion/home.html
nasal2009@umh.ac.be

15 June 2009
Balto-Slavonic Natural Language Processing (BSNLP 2009)
Cracow (Poland)
http://erssab.u-bordeaux3.fr/BSNLP

17–19 June, 2009
Phonetics and Phonology in Iberia 2009 (PaPI 2009)
Las Palmas de Gran Canaria (Spain)
http://www.congresos.ulpgc.es/papi2009/
papi2009@ulpgc.es

18–19 June, 2009
6es Journées d'Études Linguistiques (JEL'2009)
Nantes (France)
http://www.lettres.univ-nantes.fr/lling/jel2009/
olivier.crouzet@univ-nantes.fr

19 June, 2009
4th Workshop on Spanish within the Tones and Break Indices
Las Palmas de Gran Canaria (Spain)
http://www.congresos.ulpgc.es/papi2009/workshop.html
pilar.prieto@uab.cat

21–25 June 2009
13th International Conference "Speech and Computer" (SPECOM'2009)
Saint-Petersburg (Russia)
http://www.specom.nw.ru

24–26 June, 2009
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'09)
Senlis (France)
http://www-lipn.univ-paris13.fr/taln09/index.php?conf=RECITAL

June 24–26, 2009
16e Conférence sur le Traitement Automatique des Langues Naturelles (TALN'09)
Senlis (France)
www-lipn.univ-paris13.fr/taln09

5–7 July 2009
16th International Conference on Digital Signal Processing (DSP 2009)
Santorini (Greece)
http://www.dsp2009.org

6–9 July, 2009
International Association of Forensic Linguists (IAFL)
Amsterdam (The Netherlands)
http://iafl09.let.vu.nl/
iafl09@let.vu.nl

7–10 July, 2009
International Workshop on Balto-Slavic Accentology 5 (IWoBA 5)
Opava (Czech Republic)
roman.sukac@fpf.slu.cz

9–11 July, 2009
Multimodality of Communication in Children (MULTIMOD 2009)
Toulouse (France)
http://w3.eccd.univ-tlse2.fr/multimod2009/
guidetti@univ-tlse2.fr

13–16 July, 2009
2nd International Conference on Philology, Literatures and Linguistics
Athens (Greece)
www.atiner.gr/docs/Literature.htm

28–29 July, 2009
Psychocomputational Models of Human Language Acquisition (PsychoCompLA-2009)
Amsterdam (The Netherlands)
http://www.colag.cs.hunter.cuny.edu/psychocomp/

3–5 September, 2009
2nd ISCA Workshop on Speech and Language Technology in Education (SLaTE 2009)
Wroxall Abbey Estate, Warwickshire (UK)
http://www.sigslate.org

6–10 September, 2009
Interspeech 2009
Brighton (UK)
http://www.phon.ucl.ac.uk/home/interspeech2009/

9–11 September, 2009
Discourse & Prosody Interface (IDP 09)
Paris (France)
http://idp09.linguist.univ-paris-diderot.fr
idp09@linguist.jussieu.fr

10–13 September, 2009
The 8th International Conference on Auditory-Visual Speech Processing (AVSP) 2009
Norwich (UK)
http://www.avsp2009.co.uk

13–18 September, 2009
International Conference on Text, Speech and Dialogue (TSD 2009)
Plzeň (Czech Republic)
http://www.tsdconference.org/

14–16 September, 2009
Recent Advances in Natural Language Processing (RANLP-09)
Borovets (Bulgaria)
http://www.lml.bas.bg/ranlp2009

17–18 September, 2009
Workshop on Prosody and Meaning (WPM)
Barcelona (Spain)
http://prosodia.uab.cat/prosodyandmeaning/home/index.php
pilar.prieto@uab.cat

24–26 September, 2009
Gesture and Speech in Interaction (GESPIN 2009)
Poznań (Poland)
gespin2009@gmail.com
http://www.ifa.amu.edu.pl/~gespin/

29 September – 1 October, 2009
19th Czech-German Workshop on Speech Processing
Prague (Czech Republic)
http://www.ufe.cz/events/workshop-2009.php

14–16 October, 2009
Translating Beyond East and West
Prague (Czech Republic)
http://utrl.ff.cuni.cz/Translation-Beyond-East-and-West/
prague@ff.cuni.cz

23 October, 2009
Searching Spontaneous Conversational Speech (SSCS 2009)
Beijing (China)
http://ict.ewi.tudelft.nl/SSCS2009/

6–8 November, 2009
Language and Technology Conference 2009 (LTC 2009)
Poznań (Poland)
vetulani@amu.edu.pl

16–18 November, 2009
8es Rencontres Jeunes Chercheurs en Parole (RJCP)
Avignon (France)
http://rjcp2009.univ-avignon.fr/

20 November, 2009
De la perception à la compréhension d'une langue étrangère
Université de Strasbourg
moritz@umb.u-strasbg.fr

4–5 December 2009
3es Journées de Phonétique Clinique
Aix-en-Provence (France)
http://www.lpl-aix.fr/~jpc3/

13–17 December, 2009
Automatic Speech Recognition and Understanding Workshop (ASRU2009)
Merano (Italy)
http://www.asru2009.org/

23–24 December, 2009
9th Conference on Language Engineering
Cairo (Egypt)
http://www.esole.org

14–16 January, 2010
Conference on the Word in Phonology
New York, NY (USA)
http://www.cunyphonologyforum.net/word.php

4–6 February, 2010
Second Language Phonology (CASPSLP2010)
Gainesville, FL (USA)
http://caspslp2010.edublogs.org/

24–26 February, 2010
Prosodic Typology – State of the Art and Future Prospects
Berlin (Germany)
http://www2.hu-berlin.de/dgfs/

15–19 March, 2010
International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2010)
Dallas (USA)
http://www.icassp2010.com/

19–21 March, 2010
Ultrafest V
New Haven, Connecticut (USA)
http://www.haskins.yale.edu/conferences/ultrafestV.html

24–25 March, 2010
2nd Belgrade International Meeting of English Phoneticians (BIMEP 2010)
Belgrade (Serbia)
bimep.2010@gmail.com

22–24 April, 2010
European dyslexia conference
Bruges (Belgium)
http://www.khbo.be/eda-khbo-dyslexiaconference

1–3 May, 2010
New Sounds 2010
Poznań (Poland)
http://ifa.amu.edu.pl/newsounds/

3–5 May, 2010
The 2nd International Workshop on Spoken Languages Technologies for Under-resourced languages (SLTU'10)
Penang (Malaysia)
http://www.mica.edu.vn/sltu-2010

11–14 May, 2010
Speech Prosody
Chicago, Illinois (USA)
http://www.isle.uiuc.edu/speechprosody2010/

19–21 May, 2010
7th Conference on Language Resources and Evaluation (LREC 2010)
La Valletta (Malta)
http://www.lrec-conf.org/lrec2010/

23 May, 2010
Third International Workshop on Emotion: Corpora for Research on Emotion and Affect
La Valletta (Malta)
http://emotion-research.net/sigs/speech-sig/emotion-workshop

24–28 May, 2010
4th International Conference on Language and Automata Theory and Applications (LATA 2010)
Trier (Germany)
http://grammars.grlmc.com/LATA2010/

25–28 May, 2010
XXVIIIes Journées d'Étude sur la Parole (JEP 2010)
Mons (Belgium)
http://w3.umh.ac.be/jep2010

8–9 June, 2010
The 7th International Workshop on Natural Language Processing and Cognitive Science
Funchal, Madeira (Portugal)
http://www.iceis.org/Workshops/nlpcs/nlpcs2010-cfp.htm

29 June – 1 July, 2010
Basal Ganglia Speech Disorders & Deep Brain Stimulation
Aix-en-Provence (France)
http://aune.lpl.univ-aix.fr/~dbsspeechsymposium2010/

23 June, 2010
13th Meeting of International Clinical Phonetics and Linguistics
Oslo (Norway)
http://www.hf.uio.no/icpla2010

1–3 July, 2010
Colloque du Réseau Français de Phonologie. In memoriam Nick Clements
Orléans (France)
http://forum.bdp3.com/appels-a-com-f23/appel-a-com-colloque-du-reseau-francais-de-phonologie-t813.htm

4–8 July, 2010
International Conference on Conversation Analysis 2010 (ICCA10)
Mannheim (Germany)
http://www.icca10.org/start/

6–9 July, 2010
17th Workshop on Logic, Language, Information and Computation (WoLLIC 2010)
Brasília (Brazil)
http://wollic.org/wollic2010/instructions.html

8–10 July, 2010
12th Conference on Laboratory Phonology (LabPhon 12)
Albuquerque, NM (USA)
http://www.unm.edu/~labfon12/
labfon12@unm.edu

12–15 July, 2010
2nd World Congress of French Linguistics (CMLF-2010)
New Orleans (USA)
http://www.ilf.cnrs.fr/

12–15 July 2010
3rd Annual International Conference on Philology, Literatures and Linguistics
Athens (Greece)
www.atiner.gr/docs/Literature.htm

19–22 July, 2010
17e Conférence sur le Traitement Automatique des Langues Naturelles (TALN'10)
Montreal (Canada)
http://www.groupes.polymtl.ca/taln2010

6–10 September 2010
11th International Conference on Text, Speech and Dialogue (TSD 2010)
Brno (Czech Republic)
http://www.tsdconference.org/

8–10 September, 2010
Phonlex 2010 International Conference
Toulouse (France)
http://phonlex.free.fr

16–18 September, 2010
Language Teaching in Increasingly Multilingual Environments: From Research to Practice
Warsaw (Poland)
http://www.ils.uw.edu.pl/ltime.html

20–22 September, 2010
20th Czech-German Workshop on Speech Processing
Prague (Czech Republic)
http://www.ufe.cz/events/workshop-2009.php

23–25 September, 2010
Laboratory Approaches to Romance Phonology
Provo, Utah (USA)
http://hispling.byu.edu/larp/

26–30 September, 2010
Interspeech 2010
Makuhari (Japan)
http://www.interspeech2010.org/

27 September – 1 October, 2010
Summer School "Cognitive and Physical Models of Speech Production, Speech Perception and Production-Perception Interaction"
Berlin (Germany)
http://summerschool2010.danielpape.info/

8–10 October, 2010
Phonetics Today
Moscow (Russia)
http://phonetics.rli.ru/en

11–16 October, 2010
XXXIIIe Colloque international de linguistique fonctionnelle
Korfu (Greece)
www.dflti.ionio.gr/silf2010

29–30 October, 2010
Phonetic Universals
Leipzig, Germany
http://www.eva.mpg.de/lingua/conference/10-PhoneticUniversals/index.html

27–28 November, 2010
ICAL 2010 – The 3rd International Conference on Applied Linguistics
Chiayi (Taiwan)

2011

6–9 January, 2011
Panel in MLA Discussion Group- Teaching Phonetics and Phonology
Los Angeles (USA)
http://www.mla.org

12–14 January, 2011
CUNY Conference on the Phonology of Endangered Language
New York City, NY (USA)
http://cunyphonologyforum.net/endan.php

19–21 May, 2011
Quatrièmes Journées de Phonétique Clinique
Strasbourg (France)
http://misha1.u-strasbg.fr/IPS/

27–31 August, 2011
Interspeech 2011
Florence (Italy)
http://www.interspeech2011.org/IS2011-(Welcome).html

CALL FOR PAPERS

The *Phonetician* will publish peer-reviewed papers and short articles in all areas of speech science including articulatory, acoustic phonetics, speech production and perception, speech synthesis, speech technology, applied phonetics, psycholinguistics, sociophonetics, history of phonetics, etc. Contributions should primarily focus on experimental work but theoretical and methodological papers will also be considered. Papers should be original works that have not been published and are not considered for publication elsewhere.

Authors should follow the guidelines of the *Journal of Phonetics* for the preparation of their manuscripts. Manuscripts will be reviewed anonymously by two expeerts of the field. The title page should include the authors' names and affiliations, address, e-mail, telephone, and fax numbers. Manuscripts should include an abstract of no more than 150 words and up to four keywords. The final version of the manuscript should be sent both in .doc and in .pdf files. It is the authors' responsibility to obtain written permission to reproduce copyright material.

All kinds of manuscripts should be sent in electronic form (.doc and .pdf) to Prof. Dr. Mária Gósy (e-mail: gosy@nytud.hu).

We encourage are colleagues to send manuscripts for our newly released section entitled MA research: Introduction. MA students are invited to sum up their research in the area of phonetics answering the questions of motivation, topic, goal, and results (no more than 1,200 words).

INSTRUCTIONS FOR BOOK REVIEWERS

Reviews in The *Phonetician* are dedicated to books related to phonetics and phonology. Usually the editor contacts prospective reviewers. Readers who wish to review a book mentioned in the list of "Publications Received" or any other book, should address the editor about it.

A review should begin with the author's surname and name, publication date, the book title and subtitle, publication place, publishers, ISBN numbers, price, page numbers, and other relevant information such as number of indexes, tables, or figures. The reviewer's name, surname, and address should follow "Reviewed by" in a new line.

The review should be factual and descriptive rather than interpretive, unless reviewers can relate a theory or other information to the book which could benefit our readers. Review length usually ranges between 700 and 2500 words. All reviews should be sent in electronic form to Prof. Dr. Judith Rosenhouse (e-mail: swantech@013.net.il).

ISPhS MEMBERSHIP APPLICATION FORM

Please mail the completed form to:

Secretary General:
Prof. Dr. Mária Gósy, DSc
Secretary General's Office:
Kempelen Farkas Speech Research Laboratory
Hungarian Academy of Sciences
Benczúr u. 33
H-1068 Budapest
Hungary

I wish to become a member of the International Society of Phonetic Sciences

Title: _____ Last Name: _____ First Name: _____

Company/Institution: _____

Full mailing address: _____

_____

Phone: _____ Fax: _____

E-mail: _____

Education degrees: _____

Area(s) of interest: _____

The Membership Fee Schedule (check one):

1. Members (Officers, Fellows, Regular)	$ 30.00 per year
2. Student Members	$ 10.000 per year
3. Emeritus Members	NO CHARGE
4. Affiliate (Corporate) Members	$ 60.000 per year
5. Libraries (plus overseas airmail postage)	$ 32.000 per year
6. Sustaining Members	$ 75.000 per year
7. Sponsors	$ 150.000 per year
8. Patrons	$ 300.000 per year
9. Institutional/Instructional Members	$ 750.000 per year

Please charge my credit card VISA / MASTER CARD (circle one)

Owner (PLEASE PRINT):

Credit Card: # _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ Exp. date: _ _ / _ _

□ I have enclosed a cheque (in US $ only), made payable to ISPhS.

Date _____ Full Signature _____

□ Students should provide a copy of their student card.

NEWS ON DUES

Dues: Your dues should be paid as soon as it convenient for you to do so. Please note: dues are to be sent either to your Regional Secretary or directly to our treasurer.

Treasurer: Unless you pay your dues to one of the Regional Secretaries, please send them directly to Professor Ruth Huntley Bahr, Ph.D., Dept. Comm. Sci. & Dis.,
4202 E. Fowler Ave., PCD 1017, Univ. South Florida, Tampa, FL 33620-8200 USA,
Tel.: +1.813.974.3182, Fax: '1.813.974.0822,
e-mail: rbahr@chumail.cas.usf.edu

VISA and MASTERCARD: You now have the option to pay your ISPhS membership dues by credit card if you hold a VISA or MASTERCARD and pay directly to the Treasurer. Dues paying members have by now received their yearly statement which contains an authorization slip for VISA or MC. If you do not have a copy, please contact the Treasurer and she will send you one.

The Fee Schedule:

1. Members (Officers, Fellows, Regular)	$ 30.00 per year
2. Student Members	$ 10.000 per year
3. Emeritus Members	NO CHARGE
4. Affiliate (Corporate) Members	$ 60.000 per year
5. Libraries (plus overseas airmail postage)	$ 32.000 per year
6. Sustaining Members	$ 75.000 per year
7. Sponsors	$ 150.000 per year
8. Patrons	$ 300.000 per year
9. Institutional/Instructional Members	$ 750.000 per year

Special members (categories 6–9) will receive certificates; Patrons and Institutional members will receive plaques, and Affilate members will be permitted to appoint/elect members to the Council of Representatives (two each national groups; one each for other organizations).

Libraries: Please encourage your library to subscribe to *The Phonetician*. Library subscriptions are quite modest – and they aid us in funding our mailings to phoneticians in Third World Countries.

Life members: Based on the request of several members, the Boars of Directors has approved the following rates for **Life Membership** in ISPhS:

Age 60 or older:	$ 150.00
Age 50–60:	$ 250.00
Younger than 50 years:	$ 450.00

Regional Secretaries: You can now send your fees directly to your Regional Secretary if you have one – and, best yet, you may use local currency! Please note, however, that the Regional Secretaries are not yet up to accept credit card payments.