A CALL SYSTEM FOR CORRECTING VOWEL INSERTIONS IN ENGLISH SPOKEN BY NATIVE SPEAKERS OF JAPANESE

Goh Kawai, Ching Siu Lim, and Keikichi Hirose
University of Tokyo, Department of Information and Communication Engineering
goh@kawai.com, chingsiu@gavo.t.u-tokyo.ac.jp, hirose@gavo.t.u-tokyo.ac.jp
http://www.gavo.t.u-tokyo.ac.jp/

ABSTRACT

We developed a computer-aided language learning system for correcting vowel insertion errors in English pronounced by native speakers of Japanese learning English as a foreign language. The system's core component is a speech recognizer using a pronunciation lattice of American English phones plus Japanese vowels where anaptyxis may occur. The system displays English words and sentences on the computer screen and asks the learner to read them aloud. The reading material contains consonant clusters and syllable-final consonants that trigger vowel insertion errors. Loans from English having fixed Japanese pronunciations are included to illustrate pronunciation differences between the two languages. Phonological rules convert pronunciation patterns of correct English to Japanese-accented English. The system alerts the learner whenever inserted vowels are detected. The learner can adjust the speech recognizer's sensitivity to practice at different levels of difficulty.

1. INTRODUCTION

Native speakers of Japanese have difficulty pronouncing English consonant clusters because the syllabic structure of Japanese is more restricted than that of English. Japanese has a maximum of two syllable-initial consonants and one syllable-final consonant, while English allows more consonant combinations. Loan words are an example of Japanese phonotactics carrying over to English. Loans invariably undergo anaptyxis or proparalepsis [1][2][3][4].

Inserting vowels within consonants clusters or after syllable-final consonants is a prevalent error in English spoken by Japanese. Anaptyxis mutilates the syllable and stress structure of English, and anaptyctic speech is incomprehensible to native speakers of English even after considerable exposure to Japanese-accented speech. However most Japanese teachers of English overlook anaptyxis because they understand anaptyctic speech perfectly and are unaware of the severe impact anaptyxis has on intelligibility.

With this problem in mind, we implemented a system for automatically detecting inserted vowels in Japanese-accented English. The system identifies where vowels were inserted and instructs learners how to pronounce the target utterance correctly. The remainder of this paper describes the system's structure and evaluation experiment results.

2. SYSTEM STRUCTURE

The system's core is a speech recognizer (HTK [6]) running in forced alignment mode (i.e., phone labels are obtained given a correct transcription of the utterance or otherwise tightly constrained language model). The speech recognizer uses both English and Japanese monophone HMMs. Monophone HMMs for Japanese and English are trained separately on language-dependent native-speaker speech data as in regular monolingual speech recognition. The two HMM sets are used together during the recognition phase so that English phones and Japanese phones that can be substituted for English phones are both allowed. In order to combine the monophone sets of two languages, the set of features used in the HMMs must be identical (we use tied-mixture continuous-density monophone models with 12th-order melcepstra, their deltas and delta-deltas, and delta and delta-delta power). In addition, the training data's acoustic characteristics (sampling frequency, number of sampling bits, level of background noise, frequency response of microphone, and so forth) should match as closely as possible.

Implementing our method is straightforward because it uses only native speech of English and Japanese to train acoustic models. Training HMMs on non-native speech is not necessarily practical for two reasons: first, building non-native corpora in magnitudes comparable to existing native corpora is a major undertaking, and second, we probably need more data than existing native corpora because non-native speech probably has wider variance than native speech (a reasonable assumption because by definition non-natives span the range between nativeness and total non-nativeness).

The pronunciation lattice consists of phones that appear in the correct pronunciation (these are referred to as "obligatory phones" because they must appear in the learner's pronunciation), plus vowels that might be inserted (these are called "anaptyctic vowels"). The speech recognizer always detects obligatory phones. Anaptyctic vowels are detected if found in the speech signal. Anaptyctic vowels are paired with obligatory phones modeled as null phones.

Phonological rules are used to generate possible Japanese-accented pronunciations lattices of English words. The vowel [o] is inserted after [t] or [d] (e.g. [tore:] "tray"), [i] after alveolar affricates (e.g. [ri:chi] "reach"), and [u] otherwise (e.g. [suta:] "star"). Along with anaptyxis, gemination occurs at syllable-final stops following short vowels (e.g. [beddo] "bed", [pussh:u] "push"). Figure 1 shows the pronunciation lattice for the word "speech" including obligatory phones and anaptyctic vowels.

The system displays English words, phrases, or sentences on the computer screen and instructs the learner to read them aloud. The reading material consists of English words containing consonant clusters and syllable-final consonants that trigger vowel insertion. Many of the words exist as loans, making it likely that learners will mispronounce them. Table 1 shows a partial list of words and phrases trained by the system. Figure 2 shows the process flow of the system. Figure 3 shows an example of the feedback display. The system alerts the learner whenever anaptyctic vowels are recognized.

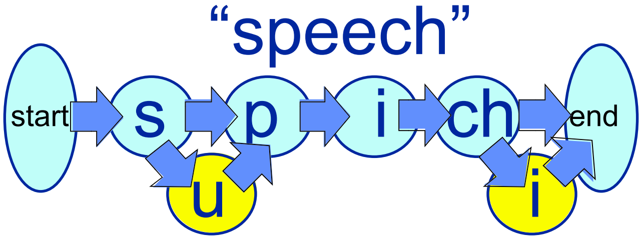


Figure 1. Example of pronunciation lattice including obligatory phones and anaptyctic vowels for the word "speech". Obligatory phones are shown in the main path, and anaptyctic vowels in alternate paths.

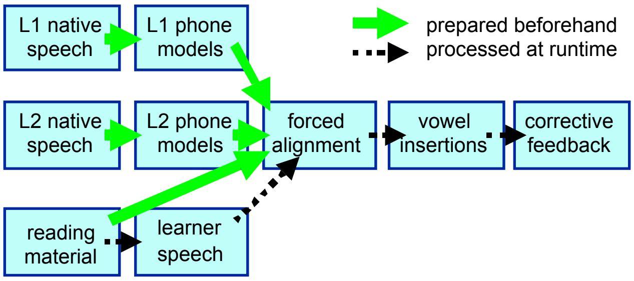


Figure 2. Process flow of the system. Solid lines show processes that are prepared ahead of time. Dashed lines indicate processes that happen when the learner is present.

Table 1. Examples of words and phrases trained by the system. Words containing consonant clusters that are disallowed by Japanese phonotactics are chosen, with heavy concetration words that exist as loans.

Isolated words	•extra, •touch •train
Loan words in carrier phrases (only words in [brackets] are graded)	•I have an [atlas] and [album] at home. •I have an [evening dress] and [turtle neck sweater] at home.
Sentences	•The trains were filmed in the Alps. •Please pay promptly

3. EVALUATION EXPERIMENT

We ran evaluation experiments to verify the performance of the component technology. 19 native speakers of Japanese (16 male, 3 female), all University of Tokyo undergraduate students with no prior experience with the system, used the system and read all words and phrases once each. Of the 19 subjects, 16 subjects (14 male, 2 female) used a close-talking noise-cancelling microphone (Sennheiser HMD-25) in a fairly noisy computer terminal room; there were multiple conversations happening in the vicinity of the subjects. The remaining 3 subjects (2 male, 1 female) used a desktop microphone (Sony ECM-K8 electret condenser in high-gain, cardiod-directivity mode) in the same computer room; the microphone was placed under the computer monitor where noises from the computer's fan and harddisk were audible. A native speaker of English determined where anaptyxis occurred via visual and audio inspection of all recordings.

Figure 4 shows scatter plots comparing human judgements with system-generated scores. Speech recorded with the desktop microphone was graded less reliably, which may have implications when learners study in groups because sharing head-mounted microphones can be cumbersome. Close-talking microphones raised the correlation between human and system scores to over 0.9. Figure 5 and table 2 compare human judgements with system scores obtained using various pruning thresholds for the speech recognizer. Results show that the system's sensitivity of detecting vowel insertions can be adjusted so that learners can practice at different levels of difficulty. For instance, pruning thresholds can be set so that the system detects more vowel insertions than human judgements or vice versa.

4. CONCLUSION

The performance of the system's component technology was verified. The next step is evaluating how effectively learners learn pronunciation skills using the system. The system can be improved by measuring the duration of fricatives that become moraic obstruents (e.g. [pusshu] "push"), resulting in a 3-mora pronunciation of a 1-mora word.

ACKNOWLEDGMENTS

We thank Kazuya Takeda (Nagoya University) and his colleagues for allowing the use of Japanese HMMs [4]. Many thanks to Steve Young (Cambridge University) and his team for permitting use of their American English HMMs [6].

REFERENCES

[1] Enomoto, Y. “Phonetic rules for foreign loan words in Japanese.” Nagoya Junior College Journal, vol 23, pp 93-100, 1985

[2] Kobayashi, Y. “Gemination in loans from English to Japanese.” Studies in English Language and Literature, vol 1, pp 97-114, 1992

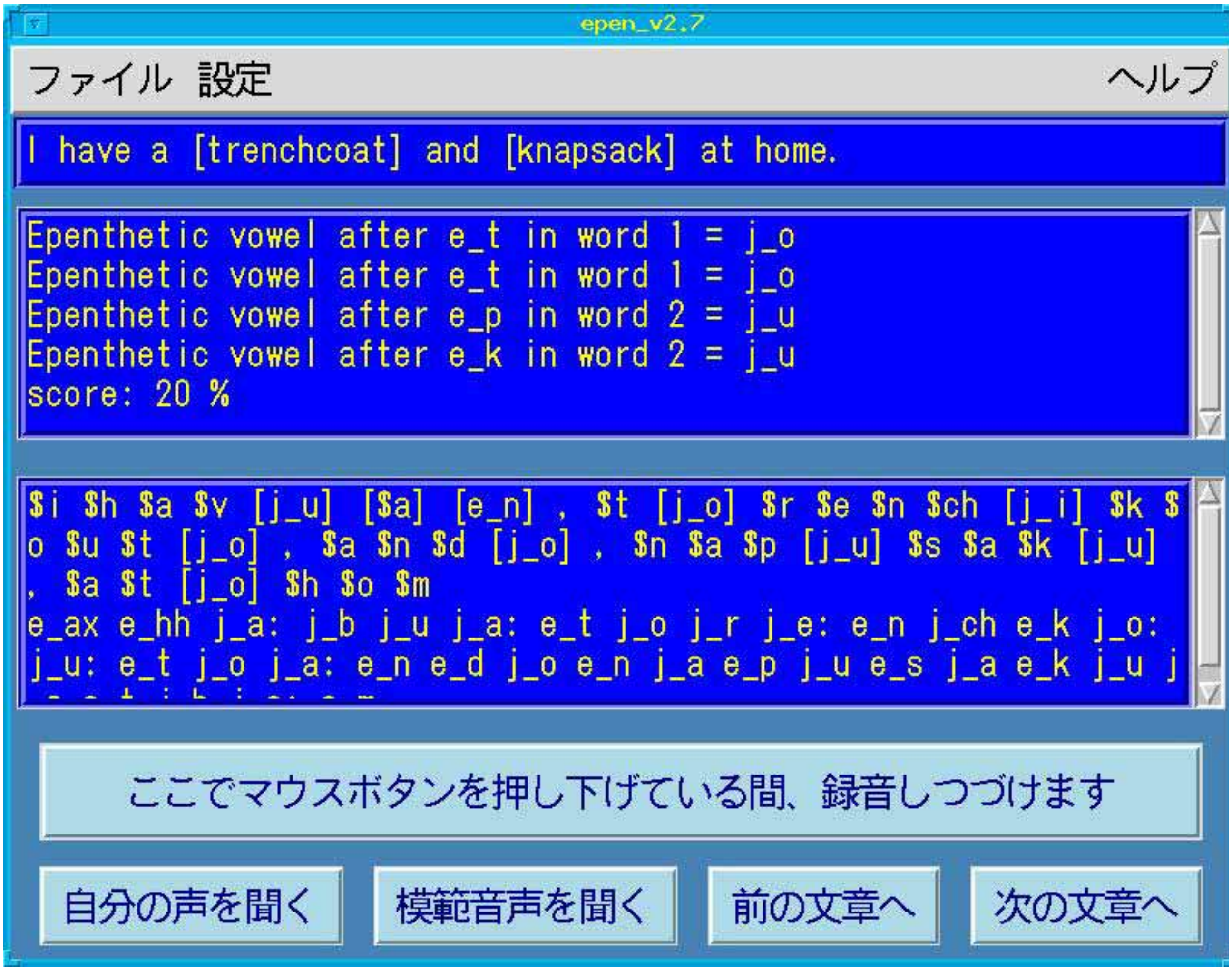
[3] Otake, T. “Consonant gemination of English loanwords in Japanese.” International Budo University Journal, vol 5, pp 101-116, 1989

[4] Takeda, K. et al “Common platform of Japanese large vocabulary continuous speech recognition research: construction of acoustic model.” IPSJ SIG Notes, paper number 97-SLP-18-3, 1997

[5] Tsuchida, O. “Automatic generation of Japanese-accented pronunciations of English words.” Unpublished senior project, University of Tokyo, 1997

[6] Woodland, P. et al “The HTK Large Vocabulary Recognition System for the 1995 ARPA H3 Task.” Proc. ARPA CSR Workshop (Arden House), 1996

[7] Young, S. et al “The HTK Book for version 2.1.1.” Cambridge University, 1997



```
I have a [trenchcoat] and [knapsack] at home.
```

```
Epenthetic vowel after e_t in word 1 = j_o
Epenthetic vowel after e_t in word 1 = j_o
Epenthetic vowel after e_p in word 2 = j_u
Epenthetic vowel after e_k in word 2 = j_u
score: 20 %
```

```
$i $h $a $v [j_u] [$a] [e_n] , $t [j_o] $r $e $n $ch [j_i] $k $
o $u $t [j_o] , $a $n $d [j_o] , $n $a $p [j_u] $s $a $k [j_u]
, $a $t [j_o] $h $o $m
e_ax e_hh j_a: j_b j_u j_a: e_t j_o j_r j_e: e_n j_ch e_k j_o:
j_u: e_t j_o j_a: e_n e_d j_o e_n j_a e_p j_u e_s j_a e_k j_u j
```

Figure 3. Sample feedback display. Learners tend to prefer detailed feedback. In this example, the learner said “trenchcoat” with a Japanese-accented [o] after the [t] s at the beginning and end of the word, and “knapsack” with a Japanese-accented [u] after [p] and [k]. As “trenchcoat” and “knapsack” have five possible locations of vowel insertions (in addition to the four the learner inserted, an [I] can be inserted after [tS] in “trenchcoat”), the system returns a 20-percent-correct score. These pieces of information are shown in the upper-half feedback window. The lower-half feedback window shows a phone-level pronunciation network (with anaptyctic vowels shown in [brackets] and English/Japanese phone combinations symbolized with $ signs), plus the actual recognized phones (English phones are shown prefixed with “e_” and Japanese phones with “j_”).

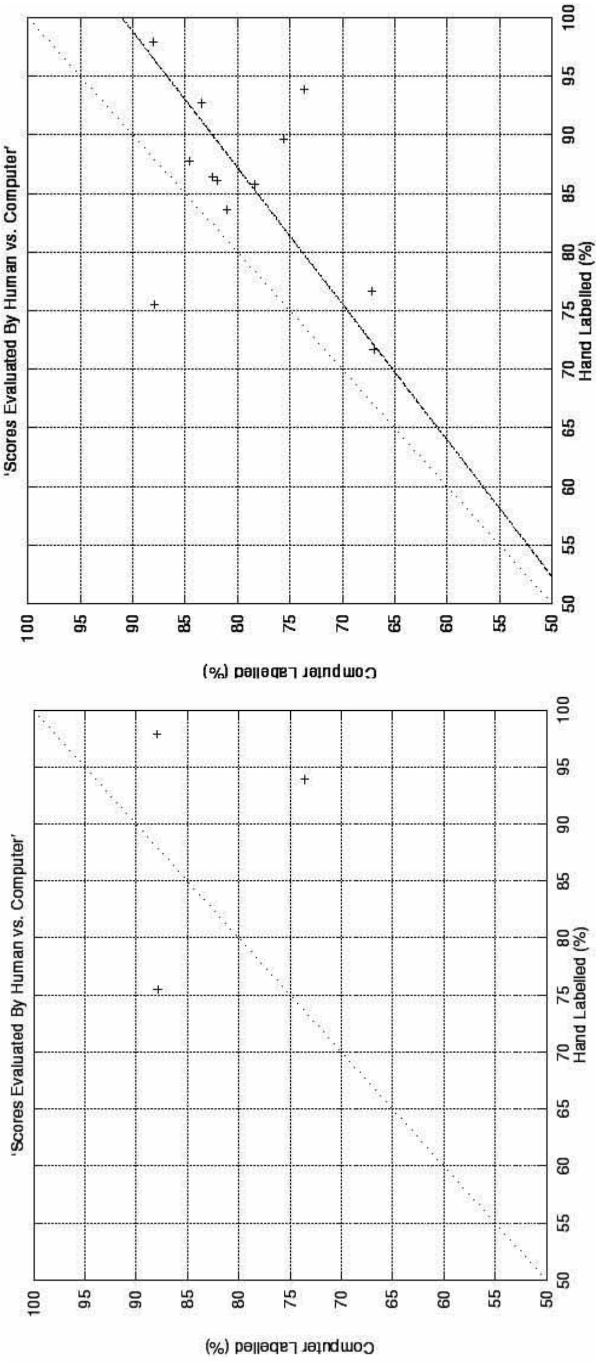


Figure 4. Comparisons of system-generated scores and human scores as a function of microphone type. The top chart is a scatter plot of scores averaged by speaker for speech data collected using a close-talking microphone (n=16). The bottom chart is data recorded on a desktop microphone (n=3). Correlation for close-talking microphone data is 0.9.

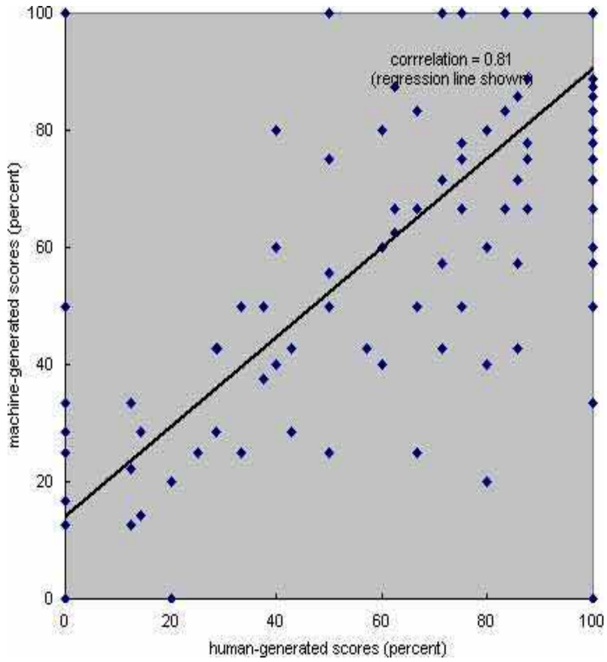


Figure 5. Best-case match of human and system-generated scores obtained by adjusting the speech recognizer's pruning threshold to 20. Correlation between human and system was 0.81. Tne regression line is overlaid. Some datapoints overlap (n=473).

Table 2. Comparison of system-generated scores and human scores as a function of speech recognizer pruning thresholds. This table shows correlations between human and system-generated scores for a range of pruning values including the one shown in Figure 5. Increasing pruning threshold allows more anaptyctic vowels to be recognized and vice versa. Adjusting pruning thresholds within reasonable limits (e.g., min –50, max 90 for r>0.7) can change the system's sensitivity towards detecting anaptyctic vowels without sacrficing reliability.

Pruning threshold	Correlation between human and system-generated scores (n=473)
-70	0.64
-50	0.70
-30	0.73
-10	0.75
10	0.75
30	0.77
50	0.75
70	0.81
90	0.73
110	0.64
130	0.50
150	0.40
170	0.30