# LANGUAGE TRAINING SYSTEM UTILIZING SPEECH MODIFICATION

*Meron Yoram* *Keikichi Hirose*

Department of Information and Communication Engineering, Faculty of Engineering
University of Tokyo, Bunkyo-ku, Tokyo 113 Japan

## ABSTRACT

In this paper, a computer assisted language training system, focusing on speech input and output, is described. The system is intended to help students of foreign language (typically Japanese or English) to improve their pronunciation, with an emphasis on prosodic features of speech.

The system incorporates a combination of speech processing techniques, in order to analyze the input speech, and to produce effective speech feedback.

The system is implemented on a Unix PC, with audio I/O capability, in a window environment.

## 1. INTRODUCTION

Computer systems intended to help students learn a foreign language, usually concentrate on writing, reading and listening skills, and communicate with the student through text input/output and limited audio output. Especially, speech input is seldom used, because of the technical difficulty to process it. Thus, these systems are far from effective in teaching correct speaking.

The system described in this paper uses speech as the main medium of interaction with the students, thus enabling the students to improve their foreign language speaking ability.

In view of researches concerning the relative weight of speech phenomena to the perception of foreign accent (defined here as any deviation from native speech)[1], the focus of this system is on teaching production of prosodic speech features, with emphasis on intonation. However, the system is also built to handle segmental features.

Rather than presenting pre-recorded audio feedback to the students, the system produces audio feedback according to the students' speech, trying to help the students hear the difference between their speech and the speech of a native speaker. To do this, several speech processing techniques are used - speech recognition, modification and synthesis, as well as analysis needed to detect significant errors in the input speech, and to decide what would be the effective way to demonstrate this to the student.

Although speech processing techniques have considerable performance limits, we believe that a well designed working scenario can enable a system to stay within these limits.

Most of the system's components are language independent. However, some components are language dependent, written especially for Japanese English and for foreign accented Japanese. Figure 1 outlines the system's structure.

Section 2 describes the methods used for recognition and segmentation of the input speech. Section 3 describes the detection of errors in the input speech. In section 4, ways of designing effective speech feedback are described, and Section 5 describes the actual speech feedback production.

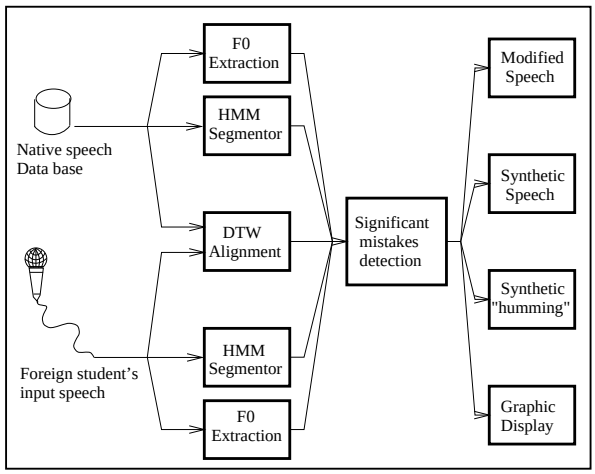


**Figure 1:** System structure

## 2. INPUT RECOGNITION

Although *full* speech recognition (with no constraints on speaker and content) produces an error rate which would not be practical, it is possible to construct a working scenario in which the content of the speech is known (with some degree of allowed variations) as in the case where the system asks

the student to repeat or read a given sentence.

## 2.1. Recognition Net

Given the speech content, it is possible to construct a recognition net with the expected phoneme sequence of the utterance, allowing for some alternative pronunciations. The allowed variations depend on the identity of L1 (the student's native language) and L2 (the foreign language), e.g. for Japanese students of English replacements of r↔l, s↔sh, etc. should be provided for.

The recognition net should restrict the search space of recognizer as much as possible, while allowing these replacements. The construction of the recognition net is done using context dependent automaton, which uses language specific rewrite rules.

## 2.2. HMM Recognizer

After the net is built, a HMM recognizer is used to locate phoneme boundaries in the input speech. The recognizer uses context independent phoneme models, each phoneme model comprising of 3 states. The features used are 12 Cepstrum coefficients, 12 $\Delta$Cepstrum coefficients, and one $\Delta$Energy coefficient. Speech is sampled at 10 kHz, and features are calculated on frames of 25.6 msec, in 10 msec intervals. The phoneme models were trained on the OGI speaker independent, telephone, multi-lingual (native) speech database.

Although restricting the system to recognize a known speaker could improve the segmentation accuracy, the system uses speaker independent phoneme models. This is because of the difficulties in collecting training speech for each specific user, and in performing speaker dependent training in the presence of speech mistakes (which can be expected for a foreign language student), but mainly because the segmentation results using speaker independent models are acceptable, even when segmenting speech of foreign accented speech, spoken in different recording conditions.

Moreover, it is possible to ask the student to repeat a sentence if the system suspects that the segmentation failed (indicated by an exceptionally bad segmentation score or "strange" segment durations), or if the student, after hearing his own speech cut into separate syllables, detects errors.

The required accuracy of the segmentation depends on the application. For intonation training, segmentation errors within half a syllable are usually acceptable, but for segmental features or duration training, higher segmentation accuracy is needed.

## 2.3. Time Alignment

The accuracy of comparison between teacher and student utterances of the same sentence can be improved by using a Dynamic Time Warping (DTW) between the two utterances. The DTW can achieve better accuracy than the HMM recognizer, because it uses a higher time resolution (single frames, as opposed to multiple frames per state) than the HMM recognizer.

A time alignment can be produced by either HMM or DTW alone, or by combining both. When segmentation is available (both for teacher and student), it can induce an alignment between the utterances. This alignment can serve as an initial alignment, which the DTW can optimize (taking advantage of the higher time flexibility of the HMM).

When segmentation is not available (e.g. in a scenario when the system asks the student to imitate an utterance, without reference to it's content), no initial alignment is used in the DTW process (which is then more vulnerable to extreme stretching and pause insertion).

# 3. ERROR DETECTION

In the detection of pronunciation errors, two important considerations are:

1) In order to be effective, a language training system should help the students to concentrate on specific points in their speech, one at a time.

2) Students need to be told exactly where they made a mistake, what is the correct pronunciation, and if possible - what was the mistake, as they might not hear the difference, or interpret the difference wrongly. This is because students might not be able to perceive some phenomena in the language they are learning, when such phenomena do not exist in their own native language. When they encounter such a phenomenon, they tend to hear and produce it in terms of their own native language (L1 interference [1]).

As each sentence can be uttered in many ways (depending on context, personal habits, attitude etc.), and as currently no general models for deciding on the correctness of a specific utterance are known, it was decided that the student should be asked to imitate an utterance by the teacher, which is considered to be the "correct" pronunciation.

Even so, it is not possible to consider any deviation from the teacher's pronunciation as an error. Rather, only *significant* deviations should be considered as errors. Two ways of finding significant deviations are used.

## 3.1. Error Function

First, an error function is calculated - a measure of similarity between the student's utterance and the reference pronunciation.

For segmental features, the segmentor's result can be used. During the segmentation, a similarity score is calculated for each phoneme. Choices made by the segmentor between sev-

eral phoneme options can also be used to indicate errors.

For intonation, another method is used. First, the intonation of both speakers is normalized : $F0_{Norm} = \frac{F0 - F0_{Ave}}{STD_{F0}}$, where $STD_{F0}$ and $F0_{Ave}$ are the standard deviation and average of the F0 (fundamental frequency) of each speaker. this normalization is intended to remove some of the individual speaker characteristics.

Then a *heuristic* error function is calculated between the two utterances (using the time alignment), as a combination of the difference in the normalized intonation and the difference in the slope of the normalized intonation (the exact way of combining these features may be language dependent).

## 3.2. Intonation Model

Another way to detect significant intonation deviations is by using an intonation model, or intonation parameterization. In our system, this is currently used for Japanese speech only (L2 = Japanese), and is based on the the superpositional model of F0 contours [3] (henceforth- F0 model).

The F0 model enables to represent the intonation of a sentence by a small number of intonation commands (phrase and accent commands), which have perceptually meaningful parameters (timing, intensity, and duration). The advantage of using this representation is that in addition to detecting the place of errors, it can provide a framework of "high level" description of the error, telling the student which parameter should be changed, and how.

First, the F0 model parameters of the teacher's sentence are extracted. This process can be manually supervised beforehand, as the extraction process isn't perfect yet (in the case of using synthetic speech as reference, the parameters are automatically produced).

Next, using the time alignment between the utterances, an initial estimate of the F0 model parameters for the student's speech is created, by aligning the timing of the teacher's intonation commands, and copying the other parameters as they are. Then a restricted analysis by synthesis (AbS) process is performed, in which the parameters are optimized to yield a better fit to the student's intonation. In this optimization, limited insertions and deletions of intonation commands are allowed.

Last, the intonation commands of the two utterances are compared (again, using the time alignment). Interpreting the differences between the intonation commands must be done carefully. If the student's intonation is very different, the resulting intonation commands may be too different to allow a meaningful explanation of the difference (in this case, the optimization constraints may not allow parameters to converge to a good fit). When the overall intonation command structure is similar in the two utterances, a difference grater than some (adjustable) threshold is considered significant (in particular, insertion or deletion of short intonation commands reliably indicate a stressing error, even when the overall structure is quite different).

For English, a third method was tested - using intonation stylization [4]. The intonation is divided into piecewise linear segment. Then, either an error function is calculated between the stylized intonations, or each syllable is classified to one of several intonation classes (flat, fall, rise-fall, etc.) and then the classes are compared between the utterances. Although this method abstracts some redundant intonation information (which is useful for a graphic intonation display), it was not found to give significantly different indications than the heuristic error function method.

# 4. FEEDBACK DESIGN

After deciding what are the points which the student should concentrate on, effective demonstration ways are needed to help the student understand these points.

The system can give the student feedback by using a graphic display, by giving a text explanation, and by speech output.

The graphic display can draw the intonation curve (in it's original form or in several stylized ways), draw the error function along the sentence text (letting the student see where errors were made in the sentence), show the F0 model intonation commands etc.

Text output is used to explain the meaning of the difference in F0 model parameters and to show the value of the error function at each specific point or summed over the whole utterance, in order to give an indication of correctness.

However, unlike other systems, the emphasis in this system is on trying to find effective methods of using speech feedback.

Because of perception problems, simple playback of the native and students' speech as it is, might not be enough for the students to be able to understand the significant differences between the utterances.

Perception is made even more difficult when the student has to compare between utterances spoken in different voices, and when the significant featured are accompanied by many irrelevant details.

## 4.1. Exaggerated Speech

Emphasizing the significant differences is achieved by creating exaggerated speech. This tries to imitate human teachers, who often try to emphasize differences by exaggerating them. For example, if the student placed a stress on the wrong syllable in a word, the teacher would contrast the correct and the wrong stressing by exaggerating prosodic features at the point of emphasis - a more extreme pitch value, higher energy and slower speech rate.

The system tries to produce the same kind of prosodic mod-

ification. In places errors were detected, the speech rate is slowed, and the intensity increased. If an intonation model is used, the wrong parameter is made even more wrong. If a model is not used, the pitch value is made more extreme at each point (low pitch becomes lower, and high pitch becomes higher).

### 4.2. Voice Change

Another problem with speech feedback, is the choice of voice to be used for feedback. It has been argued that comparing two sentences may be easier if both were spoken by the same speaker. Adopting this approach, the system has the following options:

1) Modifying the *student's own voice*: This can help the student hear the correct intonation without distractions caused by listening to another speaker's voice.

2) Creating *synthetic humming* from the teacher and student's original (or exaggerated) speech. This can help the student by removing some content distractions. Synthetic humming is created by summing sinusoid waveforms, using a given pitch and intensity specification. For prosodic training, this can help remove some of the irrelevant details, by removing segmental data from the speech.

3) Using a *text-to-speech synthesizer* to create speech according to a specified prosody. This method, too, allows the student to hear two different utterances with the same voice, but can help to prevent the uneasiness some students feel while hearing their own voice played back.

## 5. FEEDBACK PRODUCTION

### 5.1. Synthesis

As mentioned earlier, speech synthesis can be used both for output and as the reference (teacher) speech. Two text-to-speech synthesizers (TTS) were used. For English, a public domain synthesizer based on the Klatt formant synthesizer was used. For Japanese, a formant synthesizer TTS, developed in our lab was used. Both synthesizers were modified to be able to produce speech according to specified prosodic features (fundamental frequency, rate and intensity) derived from a given natural utterance.

### 5.2. Speech Modification

The core of the system is based on the PSOLA (pitch synchronous overlap add) paradigm, which was shown to enable flexible creation and modification of high quality speech [5]. Using PSOLA, prosodic changes are easily performed.

Although the system concentrates on prosodic features teaching, some segmental modification is also needed, in order to correct some of the more "severe" segmental errors, so that they will not be repeated in the feedback and serve as bad pronunciation examples.

For unvoiced segments, simple segment replacement (inserting the teacher's segment instead of the student's segment) produces acceptable speech. However, when voiced segments are replaced in this way, the resulting speech sound like an unnatural mix of the two speakers.

To improve the quality of the speech in this case, a method of spectral envelope modification is used, which helps to preserve the student's individual voice. The spectral envelope is copied from a correct segment by a native speaker, while the student's original voice source parameters are retained.

## 6. CONCLUSION

A foreign language training system is developed, with a focus on teaching correct speech. Various speech processing techniques are used in order to allow the system to respond to the student's speech. Especially, speech modification is used to produce utterances which will help the students perceive important speech features.

The system currently runs on a 486 based Unix PC, in an X-Windows environment, with no special hardware, at acceptable execution time [6].

Further work will concentrate on the use of the intonation model in detecting and demonstrating errors.

The system's effectiveness has yet to be tested on real students, but informal listening suggests that it can help students perceive their mistakes.

## 7. REFERENCES


1. Flege J.E., "The production and perception of foreign language speech sounds", *Human communication and it's disorders - a review* 1988, ed. Harris Winitz, pp. 224-401
2. Rooney et al., "Prosodic features for automated pronunciation improvement in the SPELL system", *ICSLP 92*, pp. 413-416
3. Fujisaki H., Hirose K., "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese", *J. Acoust. Soc. Jpn.(E)5*, 4 (1984), pp. 233-242
4. Bagshaw P.C., "Automatic prosodic analysis for computer aided pronunciation teaching", *Ph.D. thesis*, Univ. of Edinburgh, 1994
5. Valbert H. et al., "voice Transformation Using PSOLA Technique", *Speech Communication 11*, 1992, pp. 175-187
6. Some examples of modified speech and program screens, `http://WWW.gavo.t.u-tokyo.ac.jp/~meron/HomePage-e.html`