January 1999, Volume 2, Number 2
pp. 62-76 online
(page numbers in PDF differ and should not be used for reference)

USING AUTOMATIC SPEECH PROCESSING FOR FOREIGN LANGUAGE PRONUNCIATION TUTORING: SOME ISSUES AND A PROTOTYPE

Maxine Eskenazi
Language Technologies Institute
Carnegie Mellon University

ABSTRACT

In this article I will discuss the possible use of automatic speech recognition (ASR) for training students to improve their accents in a foreign language. Principles of good language training as well as the limits of the use of Automatic Speech Recognition (ASR) and how to deal with them will be discussed first. I will then use the example of the Carnegie Mellon FLUENCY system to show concretely how such a system may function. Prosody training as well as phonetics will be emphasized. Finally, I will underscore the importance of having a system that adapts to the user, again using the FLUENCY system as an example.

INTRODUCTION

The growing speed and memory of commercially available computers, together with their decreasing price, are making the creation of automatic language trainers feasible. The state of the art in speech recognition systems has also progressed impressively. A well-designed interface that takes advantage of the recognizer,s strong points, compensates for its weak areas, and implements basic second language learning principles contains all the necessary components of a viable language trainer.

This paper will address the use of speech recognition in second language pronunciation training. Starting with a look at some of the basic principles of language teaching, it will describe the possibilities and limits of automatic speech recognition for L2 pronunciation, and discuss its use in an interface created at Carnegie Mellon University (CMU).

While the use of the recognizer for learning grammatical structures, vocabulary, and culture will be touched on in passing, this paper will deal mainly with pronunciation training. Below a certain level, even if grammar and vocabulary are completely correct, effective communication *cannot* take place without correct pronunciation (Celce Murcia & Goodwin, 1991) because poor phonetics and prosody can distract the listener and impede comprehension of the message.

BASIC PRINCIPLES THAT CONTRIBUTE TO SUCCESS IN PRONUNCIATION TRAINING

Of the principles usually described as contributing to the success in pronunciation training (Kenworthy, 1987), we list five that are applicable to the automatic language training situation.

- Learners must produce large quantities of sentences on their own.
- Learners must receive pertinent corrective feedback.
- Learners must hear many different native models.
- Prosody (amplitude, duration, and pitch) must be emphasized.
- Learners should feel at ease in the language learning situation.

The first four principles refer to the "external" environment of language learning while the fifth principle addresses the learners, "internal" environments, that is, their attitudes and language learning capacities (Celce Murcia & Goodwin, 1991). Whereas very young language learners produce new sounds readily, as their age increases, ease in perceiving and producing them decreases. Adult learners need to feel self-confident and motivated in order to produce new sounds without inhibition. Learners who are ill at ease have a higher risk of performing poorly, or of even completely abandoning language learning. A brief discussion of the importance of these five principles is presented below.

Learners Must Produce Large Quantities of Sentences on Their Own

Ideally students fare best in one-on-one interactive language training situations with trained language instructors. Active production of speech in such environments prepares learners to participate effectively in meaningful conversations later on. However, one-on-one tutoring with human instructors is usually too costly and impractical. In reality, most learners attend classes in which they have to share their teacher,s attention with each other. This greatly reduces the amount of time each learner spends in producing foreign language speech.

Learners Must Receive Pertinent Corrective Feedback

In natural conversations, listeners often interrupt one another when they detect errors and when their relationship to the speaker permits them to offer corrections. In these instances they may provide a correction or simply point out the error. Corrections also occur when the intended message doesn,t get across. In this case, a correctly formed message usually results from an ensuing dialogue in which meaning is negotiated.

Ideally teachers point out incorrect pronunciation at the right time, and refrain from intervening too often in order to avoid discouraging the student from speaking. They also intervene soon enough to prevent errors from being repeated several times and from becoming hard-to-break habits. The pace of correction, that is, the maximum amount of interruptions per unit of time that is tolerable, is usually adapted to fit each student,s personality. Helpful feedback implies that the type of correction offered will give students the tools to deal with other aspects of the same pronunciation problem.

Celce Murcia and Goodwin (1991) report that many teachers use a "listen and imitate" technique involving the presentation of minimal pairs such as the English *tear* and *tore*. Their research suggests that mere repetition of sounds in context is *not* an effective way for students to learn to pronounce L2 sounds or even to *hear* them. If a sound does not already belong to a student,s repertory of speech sounds, then it will be associated with the closest equivalent in the learner,s repertory. For example, if an Anglophone who is just starting to learn French hears the sound /y/ in *tu*, he or she will most probably associate it with the sound /u/ of the English *toot*, since /y/ has not yet been taught as a separate phoneme for the student.

Teaching techniques in the past have followed the principle that in order to *hear* foreign sounds, categorize them, and produce them, students must be given specific instruction on how to *articulate* them. It was believed that learners must physically experience the articulation of the sound and be able to produce it before they can hear it as a separate, significant element in the target language. In the above example, teaching students to round their lips would be more effective than repeating a minimal pair. Yet recent research by Akahane-Yamada, Tohkura, Bradlow, and Pisoni (1996) shows that perception training alone may be just as effective.

Learner Must Hear Many Different Native Speakers

The optimal situation would be to learn from many different native teachers (Celce Murcia & Goodwin, 1991) who, in turn, would obtain educational materials, such as audio and video cassettes, that would further expose the student to an even wider range of native L2 speech. In reality, however, the number of teachers that a learner is exposed to is usually quite limited due to scheduling and financial constraints.

Prosody (Amplitude, Duration, and Pitch) Must Be Emphasized

When students start to learn a new language, some time is usually devoted to learning to pronounce phones that are not present in their native language. Yet experience shows that a person with good segmental phonology who lacks correct timing and pitch will be hard to understand. Intonation is the glue that holds a message together. It indicates which words are important, disambiguates parts of sentences, and enhances the meaning with style and emotion. It follows that prosody should be taught from the beginning (Chun, 1998).

Learners Should Feel At Ease in the Language Learning Situation

When learners are forced to produce sounds that do not exist in their native language in front of their peers, they tend to lose self-confidence (Laroy, 1995). As a result, they may stop trying to acquire L2 pronunciation by relying solely on L1 sounds. According to Laroy, student confidence can be increased by correcting only when necessary, reinforcing good pronunciation, and avoiding negative feedback. Therefore, one-on-one instruction is beneficial as it allows students to practice in front of the teacher alone until they are comfortable with the newly-acquired sounds.

Adapting feedback to the amount of interruption that each student can tolerate is another way to avoid discouraging active production and to obtain better results from correction. Avoiding incorrect feedback (e.g., telling students that they were wrong when in fact they were not) is a major challenge to the use of automatic speech processing since so far only a rather small margin of error has usually been acceptable in speech applications.

In addition, a contrastive analysis of the sound systems of L1 and L2 can also help to provide pertinent articulatory hints and anticipate problems and/or errors before they actually occur (Kenworthy, 1987). Knowing that there are no lax vowels in French, teachers can focus on instructing students how to go from a tense vowel in L1 to a lax vowel in an L2, (e.g., how to go from the English *peat* to *pit*). The same holds true when teaching Anglophones to use lip rounding in French as in *doux* - *du*.

HOW CAN AUTOMATIC SPEECH PRODUCTION BE USED FOR TEACHING L2 PRONUNCIATION?

An automatic system may be used as a complement to the human teacher in pronunciation training. In this scenario the teacher must provide a positive learning atmosphere, explain the differences between the segmental and suprasegmental features of L1 and L2 while the computer takes over those aspects of pronunciation practice that correspond to the principles outlined above.

Learners Must Produce Large Quantities of Sentences on Their Own

One of the major problems in automated language training lies in the fact that students are usually assigned a passive role. For example, if they are asked to answer a question, they can either repeat the sentence they had learned, or they could read aloud one of the written choices (Bernstein, 1994; Bernstein & Franco, 1995). In both cases, the answers are ready-made with vocabulary chosen and syntax assembled. As a result, students have no practice in constructing utterances on their own. *AuraLog* (Auralog, 1995), for example, has produced an appealing language teaching system that feeds the user,s pronunciation of one of three written sentences to the recognizer. The path of the dialogue is dependent on which of these sentences is elicited. While a certain degree of realism is attained, students do not actively construct any of the utterances they produce.

Producing an utterance means putting it together at many levels, including syntactic, lexical, and phonological. Phonology alone (as in minimal pair exercises) is an unrealistic task if the end goal is the ability to participate actively in meaningful conversations. To our knowledge, current speech-interactive language tutors do not let learners freely create their own utterances because underlying speech recognizers require a high degree of predictability to perform reliably. The need to know what the student will say is due to the fact that recognition, especially on the phone level, is not yet precise enough to be able to sufficiently recognize what was said by a foreign speaker without prior knowledge of the context of the sentence. We need to deal with imperfect recognition in order to prevent the system from interrupting students to tell them that they were wrong when, in fact, they were right, and in order to not overlook errors made by a student, since these need to be corrected before they become fossilized.

The FLUENCY project at CMU (Eskenazi, 1996) has developed a solution that enables users to participate more actively than in a multiple-choice situation. Automatic speech recognition has worked in language tutors so far only because the utterances were known ahead of time (read off the screen), and fed to the recognizer with the speech signal. The system simply matched exemplars of the phones it expected (pre-stored in memory) against the incoming signal (what was actually said). However, it is still possible to predict what the students will say to satisfy the needs of the recognizer, while giving them the freedom of constructing utterances on their own. This can be done by using elicitation techniques, similar to the drills that form the basis of methods such as British Broadcasting Company (Allen, 1968) and Audio-Lingual Method (Staff of the Modern Language Materials Development Center, 1964).

Several studies have been carried out to determine whether specifically targeted speech data can be collected using elicitation techniques such as those mentioned in (Hansen, Novick, & Sutton, 1996; Isard & Eskenazi, 1991; Pean, Williams, & Eskenazi, 1993). We know that with cooperative students only one to three possible sentences are appropriate as responses to any one of the elicitation sentences in a carefully constructed exercise. Non-ambiguous visual cue elicitation techniques developed in Pean, Williams, and Eskenazi (1993) succeeded in eliciting the desired structures over 85% of the time (varying from 70% for one sentence to 100% for several of the other sentences). In vocabulary choice, a careful search for non-ambiguous target nouns and adjectives with few synonyms yielded a highly predictable answer over 95% of the time. The technique has been successfully extended from French to British English and to German. The screen is free of prompts; practice is completely oral with students doing all of the production work themselves. Figure 1 below, from a set of exercises designed for the FLUENCY project, is an example of how this elicitation technique works:

System:	When did you **meet** her? (**yesterday**) - I met her yesterday.
	When did you **find it**?
Student:	I found it yesterday.
System:	Last Thursday
Student:	I found it last Thursday.
System:	When did **they** find it?
Student:	They found it last Thursday.
System:	When did they **introduce him**?
Student:	They introduced him last Thursday.

Figure 1. Sentence structure and prosody exercise for the FLUENCY Project (bold words are focus of exercise)

This technique provides a fast-moving exercise for the students, giving them an active rather than passive role. They acquire automatic reflexes in this way that are useful later, when they need to build an utterance during a real conversation. Then they should be able to maintain a conversational tempo rather than searching for correct structures and words.

FLUENCY allows automatic alignment of the predicted text with the incoming speech signal. Once the incoming signal has been recognized, the system can either intervene immediately, breaking into the rhythm of the exercise, or wait until the end of the exercise to start corrections. In order to ensure a higher level of success in elicitation, it is preferable to wait until the end of the exercise to start corrections. However, the system allows the teacher to override this if a student,s level and personality warrant it.

Students can practice producing answers to the same elicitation sentences over and over as often as they wish. Constant availability and patience are major qualities that make the automatic system an ideal tool for practice.

Learners Must Receive Pertinent Corrective Feedback

When teachers ask whether speech recognition can be useful in a language teaching system, they often wonder whether errors can be correctly detected and whether the system can offer appropriate feedback.

Can Errors Be Detected?

As mentioned earlier, a recognition system should be capable of correcting both segmental and prosodic errors. These two types of errors will be discussed separately since they are different in nature, and since their detection and correction imply very different procedures. Phone errors are due to a difference not only in the number and nature of the phonemes in L1 and L2, but also because the acceptable pronunciation space of a given phone may differ in the two languages. Prosodic errors, on the other hand, involve pitch, duration, and intensity. These components are the same in all languages but their relative importance, meaning, and variability differ from language to language. For example, intensity variations tend to be less frequent and show less contrast in French than in Spanish.

The methods for identifying errors in phones and prosody differ as well. The speech recognizer in a "forced alignment mode" can calculate the scores for the words and the phones in the utterance. In forced alignment, the system matches the text of the incoming signal to the signal, using information about the signal/linguistic content that has already been stored in memory. Then after comparing the speaker,s recognition scores to the mean scores for native speakers for the same sentence pronounced in the same speaking style, errors can be identified and located (Bernstein & Franco, 1995). On the other hand, for prosody errors, duration can be obtained from the output of most recognizers. In rare cases, fundamental frequency may be obtained as well. In other words, when the recognizer returns the scores for phones, it can also return scores for their duration. On the other hand, intensity of the speech signal is measured before it is sent to the recognizer, just after it has been preprocessed. It is important that measures be expressed in *relative* terms--such as duration of one syllable compared to the next--since intensity, speaking rate, and pitch vary greatly from one individual to another.

Phone Error Detection

Although researchers have been cautious about using the recognizer to pinpoint phone errors, recent work in the FLUENCY project at Carnegie Mellon (Eskenazi, 1996) has shown that the recognizer can be used in this task if the context is well chosen. Ten native speakers of American English (five male and five female) were recorded along with twenty foreign speakers (one male and one female) representing French, German, Hebrew, Hindi, Italian, Korean, Mandarin, Portuguese, Russian, and Spanish. The non-native speakers had varying degrees of proficiency in English. This must be taken into account when interpreting the results since fairly fluent speakers usually make considerably fewer phonetic and prosodic errors than less proficient ones. The sentences were phonemically transcribed by the author. Expert instructors served as judges and were asked to listen to the sentences and note errors in pronunciation. The agreement between automatic detection and human judges was used as the measure of the reliability of error detection.

Figure 2 shows the recognition results for some native and non-native male speakers when their speech was processed by the Carnegie Mellon SPHINX II automatic speech recognizer in a forced alignment mode (Ravishankar, 1996).

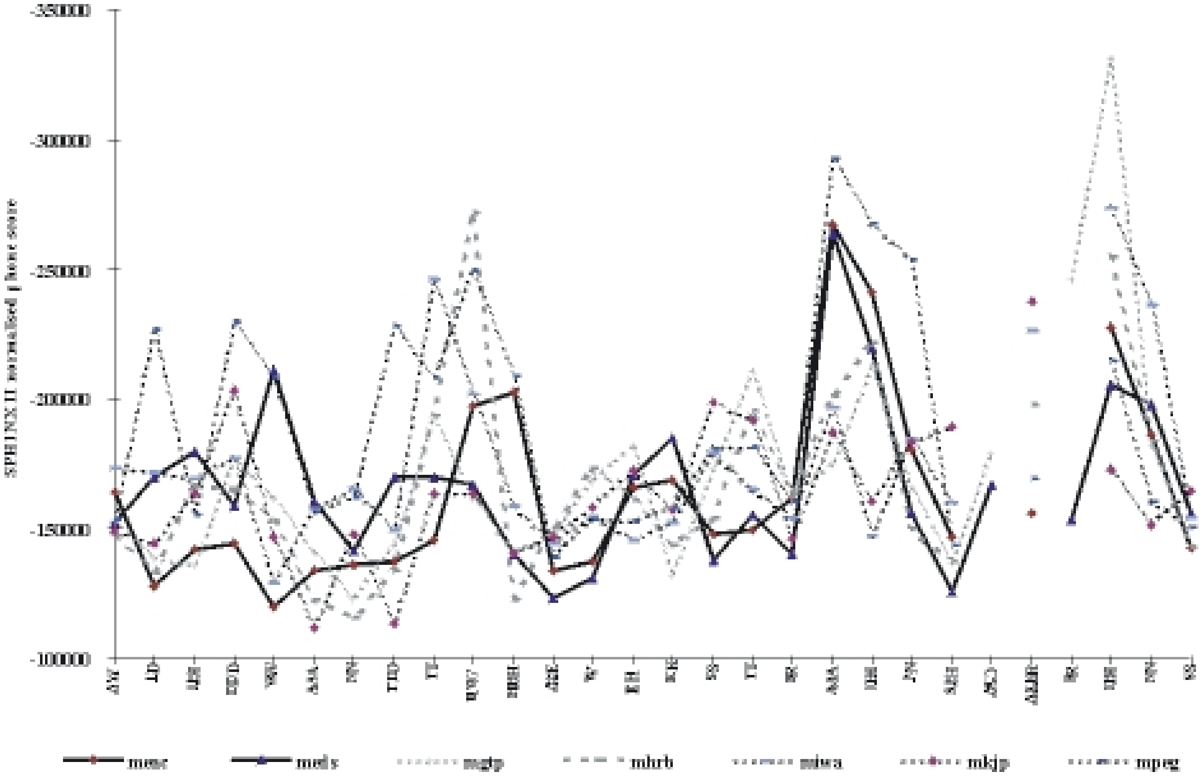


Figure 2. SPHINX II recognition of the utterance, "I did want to have extra insurance" produced by native and non-native speakers. (*x* axis = phones in the sentence)

As past experience has shown, information based on absolute thresholds of phone scores is of little use. However, for a given target phone in a given context, a non-native error can have a score that is significantly distant from the cluster of scores for native speech in the same context. In Figure 2, native speaker values are represented by solid lines while non-native values are represented by dotted lines. The first letter of the speaker reference indicates gender (M = male, F = female), while the second letter indicates the speaker,s native language (French, German, Hebrew, Hindi, Italian, Korean, Mandarin, Portuguese, Russian, Spanish.). Phones are presented on the horizontal axis. The *y*-axis represents the normalization of the phone scores given by SPHINX II over the total duration of the phone.

The following phonological variants common to native speakers of English were taken into account: for the final /t/ (labeled here /TD/) of *want* and the vocalic /r/ (labeled /AXR/ here) of *insurance*. Not all speakers show values in this region since *want* can be pronounced without the final /t/ in this context; "sur" of *insurance* can have a vocalic /r/ or a consonantal one (noted /AO/ /R/ here, according to Carnegie Mellon phone notation).

As expected, non-native speakers were not outliers all of the time, as they often produced correct phones, especially when the L2 phones were similar to phones in their L1. It can be noted that many non-native speakers did not have the /AE/ sound of /HH AE V/ (*have*) in their native language. They tended to pronounce an open /e/, (labeled /EH/ here) and sometimes /EHF/ instead of /AEV/. The German speaker was very far off in the lax vowel /IH/ of *insurance*. It is interesting to note that the female German speaker had very similar results to those of female English speakers. The human judges noted the same outliers as the automatic recognizer.

For the final /t/ of *want* (transcribed as /TD/ here), English speakers did not aspirate the stops; non-native speakers did. This contributed to perceived accent, although not enough for the tutors to note it as needing correction. This could be considered a minor deviation that contributes to the perception of a non-native accent but does not affect comprehensibility.

Therefore it appears that for this small, yet diverse population, SPHINX II confirms human judgment of incorrectly pronounced phones, independently of the speaker,s L1.

Prosody Error Detection

Prosody information is just starting to be successfully used in speech recognition in order to enhance recognition results. Fundamental frequency detection, hereinafter pitch, has the same drawback as speech recognition, namely that sufficiently correct results have yet to be obtained. Work on better pitch detectors, such as by Bagshaw, Hiller, and Jack (1993), is making the algorithms more precise within a specific application. As in CMU,s FLUENCY project, Bagshaw et al. (1993) compared the student,s contours to those of native speakers in order to assess the quality of pitch detection. Rooney, Hiller, Laver, and Jack (1992) applied this to the SPELL foreign language teaching system and attached the output to visual displays and auditory feedback. One of the basic ideas in their work was that the suprasegmental aspects of speech can be taught only if they are linked to segmental information. Pitch information includes pitch increases and decreases and pitch anchor points (i.e., centers of stressed vowels). Rhythm information shows segmental duration and acoustic features of vowel quality, predicting strong vs. weak vowels. Rooney, Hiller, Laver, and Jack (1992) also provided alternate pronunciations, including predictable cross-linguistic errors.

Tajima, Port, and Dalby (1994) and Tajima, Dalby, and Port (1996) studied timing changes and their effect on intelligibility. Using nonsense syllables (e.g., ma ma ma. . .) their approach separates segmental and suprasegmental aspects of the speech signal in order to focus exclusively on temporal pattern training.

The FLUENCY project has investigated the detection of changes in duration, amplitude, and pitch that can reliably detect where non-native speakers deviate from acceptable native values, independently of L1 and L2. Thus, if a learning system is applied to a new target language, its prosody detection algorithms do not have to be changed in any fundamental way. Since they are separate from one another, the three aspects of prosody can easily be sent to visual display mechanisms that show how to correctly produce pitch, duration, or amplitude changes as well as compare a native speaker,s production to that of a non-native speaker.

However, FLUENCY prosody training is always linked to segmental aspects, with students producing real, meaningful phones, not isolated vowels. Simple measures that compare non-native speech to native speech for ESL speakers from a variety of language backgrounds, are very promising because they detect the same errors as expert human judges.

Duration of the speech signal was first measured on the waveform, and the duration of one voiced segment was compared to the duration of the preceding segment, making the observations independent of individual variations in speed, pitch, and amplitude. A voiced segment starts at the onset of voicing after silence, or after an unvoiced consonant, and ends when voicing stops at the onset of silence or of an unvoiced consonant, independently of the number or nature of the phones it contains. Figure 3 shows the results of duration comparisons. For the sake of clarity, notations include the neighboring unvoiced consonants as well.

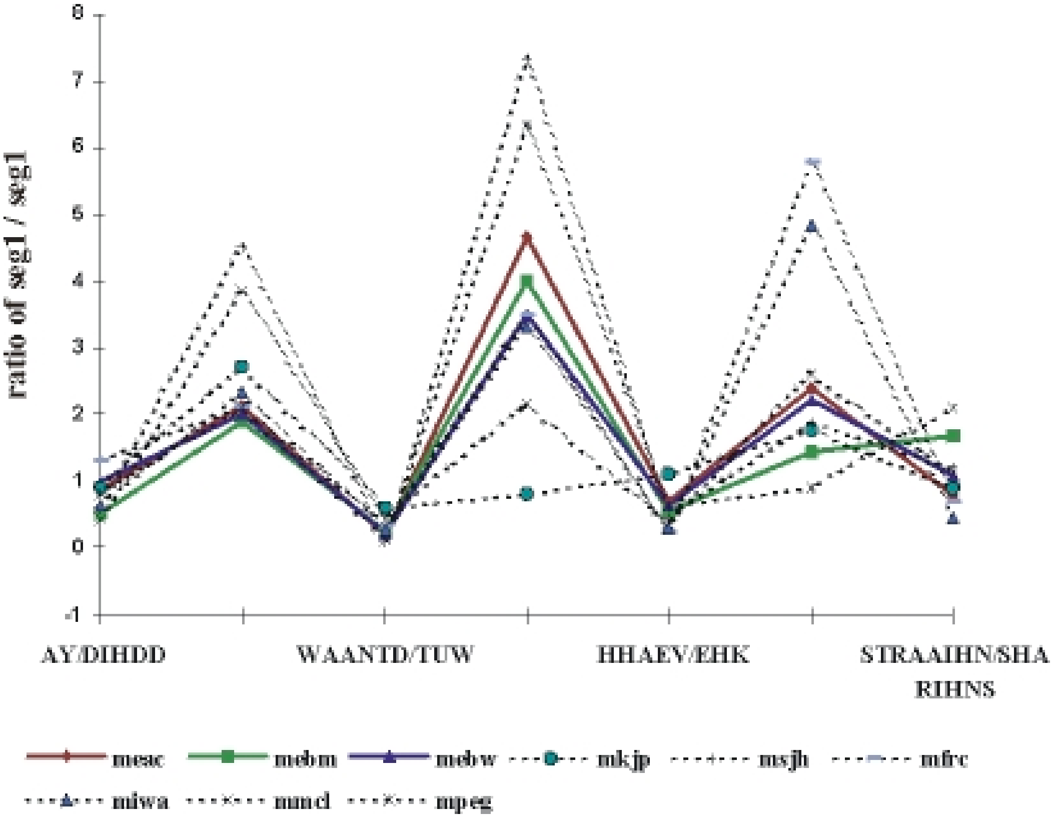


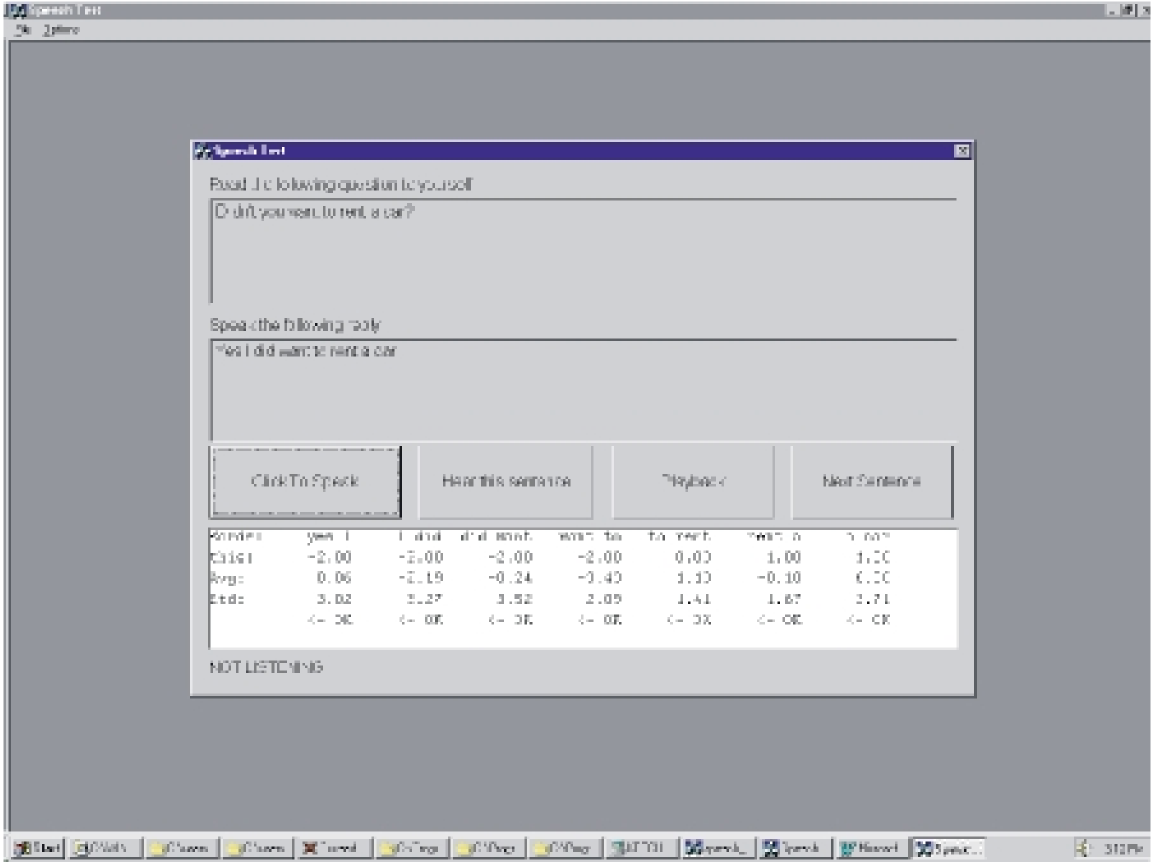
Figure 3. Two-by-two comparison of duration of voiced segments for the sentence "I did want to have extra insurance" in native and non-native speech

There are two outliers in this figure. /EHK/ (see above) is unusually long compared to the following vocalic segment for speaker mfrc and for speaker miwa. This can be due to the inability to pronounce a lax vowel since tense/lax vowel quality differences do not exist in French and Italian. Quantity differences could be learned as an easy way to first approximate the tense/lax difference.

We can also note that speaker mkjp,s *to* is of about equal length to his *have*. This is different from all of the other speakers where *have* is longer than *to*. This difference had also been noted by the expert human judges. We also examined pitch and amplitude differences between natives and non-natives in the same way (Eskenazi, 1996).

An Example of an Actual Duration Correction System

The FLUENCY system presently uses the SPHINX II recognizer to detect the student's deviations in duration compared to that of native speakers. The system begins by prompting the student to repeat a sentence. The student clicks the mouse to speak and then clicks again after he or she has finished the sentence. This is a bit cumbersome compared to using an open microphone, but it ensures more reliable detection of the speech to be compared. The speech signal and the expected text are then fed to the recognizer in forced alignment mode. The recognizer outputs the durations of the vowels in the utterance and compares them to the durations for native speakers. If they are found to be far from the native values, the system notifies the user that the segment was either too long or too short. An example of what appears on the screen can be seen in Figure 4.



Read the following question to yourself

Didn't you want to rent a car?

Speak the following reply

Yes I did want to rent a car

```
Words:    yes I    I did  did want   want to   to rent    rent a     a car
This:     -2.00    -2.00     -2.00     -2.00      0.00      1.00      1.00
Avg:       0.06    -2.19     -0.24     -3.40      1.10     -0.10      0.00
Std:       3.02     2.27      3.52      2.09      1.41      1.67      2.71
          <- OK    <- OK     <- OK     <- OK     <- OK     <- OK     <- OK
```

NOT LISTENING

Figure 4. A sample screen from the FLUENCY system

Users receive feedback about what they have just said (e.g., <-OK, <-SHORT, or <-LONG, indicating the judgment on the preceding vowel). Whole syllables are shown as opposed to individual vowels which by themselves are difficult to comprehend. Students can also hear their own speech again ("Playback"), hear the sentence pronounced by a native speaker ("Hear this sentence"), practice, and try again. In preliminary tests of the system involving 12 foreign graduate students, it took an average of three trials for a student to get all "<-OK"s. The system interface is presently being modified and adapted for phonetic error correction.

In post-trial interviews, students mentioned that they had the impression of being in charge when using the system since they could choose what function to use next and how often they wanted to use it. Most of the students in the preliminary test were reluctant to stop even when they were told that the test was over.

Learners Must Hear Many Different Native Speakers

Increased computer memory makes it possible to listen to a variety of different speakers. We now have the ability to access L2 speech from several different regions and therefore include several different accents. This is beneficial for students who are planning to travel to a specific region of a country. In addition, utterances can be repeated over and over.

Learn to Speak Spanish (Duncan, Bruno, & Rice, 1995) is one such program that attempts to expose users to a large number of speech utterances produced by a variety of speakers. In fact, videos of several different speakers pop up in the course of one exercise. The weakness of this system, however, is that while each utterance can be repeated as many times as desired, only one speaker has been recorded for a given sentence.

TOWARD AN EFFECTIVE INTERFACE

Even with a perfect recognizer, the success of the system is still dependent to a great extent on the quality of its interface. Several issues are important in this respect. First, effective correction is the basis of good interaction. Second, adapting to characteristics of the user makes the system more reassuring, thereby increasing the user's self-confidence. Finally, basing correction on the student,s past performance maximizes efficiency.

Effective Correction

Error detection is very important, but the language learning system that is limited to this function, such as parts of *TriplePlayPlus* (Syracuse Language Systems, 1994), is of marginal benefit to users who detect an error at a certain place in the sentence and then make random attempts to correct it. This self-correction may be based on their own opinion of what was wrong or, more often, on trial and error. The result can hamper true learning or improvement from taking place. The results could even be negative if students make a series of unsuccessful attempts to produce a certain sound. Without feedback, these attempts can reinforce poor pronunciation and result in fossilization.

When to Intervene

Interventions can appear to users as being either timely or irritating. Bothersome interventions tend to be caused by either recognition errors or by a system that intervenes too frequently and is too verbose. The first problem can be solved by improving the quality of the recognizer. However, at present it is common practice to restrict the way in which the recognizer is used, for example, by selecting the semantic domain or the speaking style. As far as the second problem is concerned, there are three criteria which can help decide when to intervene:

1) Would the error cause a breakdown in communication?
2) Is it a recurring or an isolated error?
3) Can the error be corrected easily? (Celce Murcia & Goodwin, 1991)

These criteria should govern the choice of the types of interventions included in a system. Naturally, errors that interfere with communication should be dealt with first.

Good Feedback

Since we do not want to simply point out errors, the method of correction is extremely important. As mentioned earlier, phone and prosody correction must be handled in two different ways. This is due to the fact that phones are completely different from one language to another whereas the elements of prosody are produced in the same way in different languages, even if they are used in different combinations. While students may need instruction on how to articulate a new phone, they do not need instruction on how to increase or decrease their pitch if they have normal hearing. They only need to be told *when* to increase and decrease it, and by how much.

Although there has been limited success in using contrasting L2 sounds in context, such as "I want a beet" with "I want a bit," the teacher can go further by accompanying the contrast drills with instructions on how to change articulator position and duration without relying on the ear (Celce Murcia & Goodwin, 1991). The idea is to get the learners to "feel" when the articulators are in place correctly. Indeed, practice with the recognizer can confirm that this has occurred. Once the phone is pronounced correctly, students can then listen to themselves. By relating what they hear to what they feel their muscles doing, they can train their ears to recognize the new sounds as distinct entities.

Phone production training can be L1-independent, going from one close cardinal vowel (most likely /a/, /i/, /u/ which have a high probability of existing in most languages), to the target vowel. A far better but more costly solution is to use an L2 vowel that has a close counterpart in L1 as a starting point. This requires knowledge of the user,s L1 as well as of differences between the L1 and L2.

In the case of prosody, a visual display is more effective than oral instructions since most of the learning results from the students, detection of where the curve representing their L2 production differs from that of a native speaker.

Adapting to the User

One way to make users feel comfortable with the system is to include options enabling them to adapt the interaction to suit their individual needs. There are many possibilities in this area. Among the options are "golden" voices and learning strategies which are both described below.

"Golden" Voices. In the FLUENCY system, the "Hear this sentence" button offers a recording of a "golden" native speaker, that is a native speaker who pronounces utterances correctly and comprehensibly and is to be imitated. Since a female speaker may have difficulty imitating a low male voice, and a male speaker may find it hard to imitate a high female voice, it is preferable to offer the user a choice of voices to pick from. However, in field tests, some users indicated that they wanted their teachers to make the choice for them.

Learning Strategies. As mentioned earlier, many training procedures have been based on showing learners how to articulate new sounds. However, recent work by Akahane-Yamada et al. (1996) has shown that new sounds can be taught by perception alone. These researchers trained Japanese speakers to hear and pronounce the /r/ - /l/ difference in English by simply listening to instructions and carefully chosen examples of minimal pairs.

This brings us to recognize that there may indeed be more than one learning strategy, and that some students may learn better by ear while others do better with visual correction (i.e., articulatory instructions on the screen). We are presently changing the FLUENCY interface to provide three feedback options: aural, visual, and a combination of both aural and visual.

However, many students may not be aware of the strategy that suits them best. To aid them in selecting a suitable strategy, we have developed a memory game made up of three parts. The first game gives the user a series of differently colored and placed buttons with corresponding tones. The buttons are played in a random series and the user must imitate the series exactly. First only one tone/button is played, then two, then three, and so forth. Each time the user repeats the series correctly, a new series, one element longer than the last, is presented. The system records how many elements are in the longest series that was correctly repeated, as well as the response times. This first game provides a base measure.

The second game presents the colored buttons without the sound. Data are recorded in the same way as for the first game. The third game presents sounds with corresponding gray buttons that constantly change position on the screen. Data are similarly recorded. We postulate that a user who does better on the third game (i.e., produces much longer correctly repeated series and/or has shorter response times) should respond better to aural training, whereas a student who does better on the second game should respond better to visual training. If a student shows no clear preference for either one, a combined method may be best.

Learning from Past Performance

The best systems are those that tally students, performance as they go along and make use of the scores in ensuing work. Consistently low scores on one type of problem (e.g., the use of an auxiliary) should then trigger the system to prompt the user to work on that problem first in the next session. Moreover, if the system detects a consistent problem with a given target phone, not only could it propose sentences designed to practice that phone or that class of phones, but it can provide feedback on the pronunciation of that phone to the exclusion of all the other phones in the sentence until better performance is achieved.

CONCLUSION

As we have seen, use of computers, and especially automatic speech processing, brings foreign language pronunciation training a wealth of new possibilities. The advantages of memory space for expanded exposure to large quantities of speech from many speakers and for multimedia corrective feedback were discussed. The use of automatic speech processing for error detection was also described. In essence, if students can be guided to use the computer as a complement to classroom instruction, the increase in practice time can help to more closely approach the advantages of total immersion learning. The ease with which the students used the computers during our tests at Carnegie Mellon is a positive step in this direction.

Despite all of this, there is still much to be done. Teachers and computer scientists need to collaborate more closely in order to make these tools more powerful and user-friendly, and develop more teaching techniques to build on the advantages of this new medium.

ABOUT THE AUTHOR

Maxine Eskenazi obtained her Doctorate in Computer Science from the University of Paris in 1984. She has worked in the field of automatic speech processing as a chargée de recherche at LIMSI-CNRS, and is now serving as a Systems Scientist at Carnegie Mellon University. She has been licensed to teach French as a foreign language by the state of Pennsylvania and has extensively taught French and English as foreign languages.

E-mail: max@parle.speech.cs.cmu.edu

REFERENCES

Akhane-Yamada, R., Tohkura, Y., Bradlow, A., & Pisoni, D. (1996). Does training in speech perception modify speech production? *Proceedings of the International Conference on Spoken Language Processing*. Philadelphia, PA.

Allen, W. S. (1968). *Walter and Connie, Parts 1-3*. London: British Broadcasting Corporation.

Auralog. (1995). *AURA-LANG User Manual*. Paris: Auralog.

Bagshaw, P., Hiller, S., & Jack, M. (1993). Computer aided intonation teaching. *Proceedings of Eurospeech , 93* , 1003-1006.

Bernstein, J. (1994). Speech recognition in language education. *Proceedings of the CALICO '94 Symposium*, 37-41.

Bernstein, J., & Franco, H. (1995). Speech recognition by computer. In N. Lass (Ed.), *Principles of experimental phonetics* (pp. 408-434). New York: Mosby.

Bowen, J. D. (1975). *Patterns of English pronunciation*. London: Newbury House.

Celce Murcia, M., & Goodwin, J. (1991). Teaching pronunciation. In Celce Murcia (Ed.), *Teaching English as a second language*. New York: Heinle and Heinle.

Chun, D. (1998). Signal analysis software for teaching discourse intonation. *Language Learning & Technology*, *2*(1), 61-77. Retrieved January 22, 1999 from the World Wide Web: http://polyglot.cal.msu.edu/llt/vol2num1/article4/index.html.

Duncan, C., Bruno, C., & Rice, M. (1995). *Learn to speak Spanish: Text and workbook*. San Francisco: Hyperglot Software.

Eskenazi, M. (1992). Changing speech styles, speakers, strategies in read speech and careful and casual spontaneous speech. *Proceedings of the International Conference on Spoken Language Processing*, Banff, Canada.

Eskenazi, M.. (1996). Detection of foreign speakers, pronunciation errors for second language training: Preliminary results. *Proceedings of the International Conference on Spoken Language Processing*. Philadelphia, PA.

Hansen, B., Novick, D., & Sutton, S. (1996). Systematic design of spoken prompts. *Proceedings of CHI,96* , 157-164.

Isard, A., & Eskenazi, M. (1991) Characterizing the change from casual to careful style in spontaneous speech. *Journal of the Acoustical Society of America*, *89* (4, Pt. 2).

Kenworthy, J. (1987). *Teaching English pronunciation*. New York: Longman.

Laroy, C. (1995). *Pronunciation*. New York: Oxford University Press.

Micro Video Corporation. (1989). *Getting started with Video Voice: A follow-along tutorial* [Brochure]. Ann Arbor, MI: Author.

Pean, V., Williams, S., & Eskenazi, M. (1993). The design and recording of ICY: A corpus for the study of intraspeaker variability and the characterization of speaking styles. *Proceedings of Eurospeech ,93* , 627-630.

Ravishankar, M. (1996). *Efficient algorithms for speech recognition* (Technical Report CMU-CS-96-143). Pittsburgh, PA: Carnegie Mellon University.

Rooney, E., Hiller, S., Laver, J., & Jack, M. (1992). Prosodic features for automated pronunciation improvement in the SPELL system. *Proceedings of the International Conference on Spoken Language Processing* (pp. 413-416). Banff, Canada.

Staff of the Modern Language Materials Development Center. (1964). *French 8, Audio-Lingual Materials*. New York: Harcourt, Brace and World.

Syracuse Language Systems. (1994). *TriplePlayPlus! User's Manual*. New York: Random House.

Tajima, K., Dalby, J., & Port, R. (1996). Foreign-accented rhythm and prosody in reiterant speech. *Journal of the Acoustical Society of America, 99*, 2493.

Tajima, K., Port, R., & Dalby, J. (1994). Influence of timing on intelligibility of foreign-accented English. *Journal of the Acoustical Society of America*, *94*, (Paper 5pSP2).

Vallette, R. M. (1967). *Modern language testing: A handbook*. New York: Harcourt, Brace and World.

Wyatt, D. (1988). Applying pedagogical principles to CALL. In W. F. Smith (Ed.), *Modern media in foreign language education* (pp. 85-98). New York: National Textbook Company.

Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. Language Learning and Technology, 2(2), 62-76. Available from: http://llt.msu.edu/.Â CALL dimensions: Options and issues in computer-assisted language learning. Mahwah, NJ: Lawrence Erlbaum Associates. Liddell, P. & Garrett, N. (2004).