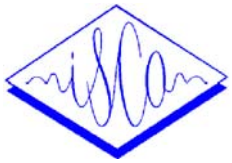

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

USER ADAPTATION IN THE FLUENCY PRONUNCIATION TRAINER

Maxine Eskenazi, Scott Hansma, John Corwin, Jordi Albornoz
Language Technologies Institute
Carnegie Mellon University
4910 Forbes Ave., Pittsburgh, Pa. 15213 USA
max@cs.cmu.edu
http://www.speech.cs.cmu.edu

ABSTRACT

Adaptation to the user is of great importance in a language training system since the user, making unfamiliar, socially unacceptable sounds, must be made to feel self-confident in order to obtain best results. The Fluency project aims at creating a pronunciation trainer for foreign language learning. It uses automatic speech recognition (CMU's SPHINX II) and is guided by basic principles in foreign language learning research. We describe efforts in this project to make the interface adaptable concerning learning strategies and "golden voices"..

Keywords: foreign language learning, adaptability

1. INTRODUCTION

The Fluency project is creating a pronunciation trainer for foreign language learning. It uses automatic speech recognition (CMU's SPHINX II [7]) and is guided by basic principles in foreign language learning research. The goal is to offer correction of both phonetic and prosodic errors, providing user-adapted interfaces.

This paper presents work done to make the system user-adaptative and to present exercises in a way that makes the user feel comfortable and "in charge". These two aspects are used by our system to give the user a feeling of self-confidence, since self confidence is a key to continued success in speaking a foreign language [3].

Techniques presently accepted as confidence bolsters [5] consist of correcting only when necessary, reinforcing good pronunciation, and avoiding negative feedback. Avoiding incorrect feedback (for example, saying a student was wrong when he wasn't) is a major challenge to the use of automatic speech processing since a small margin of error has usually been acceptable in speech applications so far.

Fluency does not judge the user. Rather, the system pinpoints specific items to be worked on. Although scores appear on the screen for development purposes, no scores will be shown to the user in future versions of the system.

While use of the recognizer for learning grammatical structure, vocabulary and culture is of equal importance, this is not the object of Fluency; exercises involving these language levels are only used as vehicles for pronunciation training. The reader is referred to the VILTS system [6] for more information about training other levels of foreign language mastery.

Pronunciation training is important; below a certain level, even if grammar and vocabulary are completely correct, communication *cannot* take place without correct pronunciation [3]. Poor phonetics and prosody distract the listener and impede comprehension of the message.

Fluency points out errors *only* where it can also provide high quality feedback, giving information to the user as to where something was not quite right and how to correct it.

2. THE BASIC TRAINING MODULES

In past work [4], we have shown it possible to use the recognizer to pinpoint errors on a smaller scale than word or sentence level. We showed that we can determine phone, duration, intensity, and pitch variations compared to a group of native speakers. The first part of the Fluency system we designed took advantage of our findings about duration. We chose duration first for three reasons: 1) we believe that prosody must be presented on the same level as phonetics; 2) the corrective feedback for duration errors necessitates a much less elaborate implementation since we do not have to show articulator placement, for example; 3) it is by far the most sure measure we have - we wanted to start with a system finctioning with the lowest error rate as possible, to establish student

confidence and serve as a basis for future work. We have since progressed to phone correction. After describing our duration module, we will discuss aspects concerning of phone correction.

2.1 Duration correction

In this module, the speaker is asked to say a sentence. The sentence will be *elicited* from the speaker by the system. At present, there is a base sentence in the box at the top of the screen as seen in Figure 1. The student then responds to it, much the way he would in an exercise in class, by saying the sentence in the second box from the top. These two boxes are being replaced by a "talking head" video in the next version of the module.
The user clicks on the "click to speak" button (leftmost of the four buttons in a row) to say his sentence. The "closed mike" is being replaced by a more user-friendly option. The first version of the system used this option in order to eliminate the marginal error associated with silence detection.
After saying his sentence, scores and arrows to where the errors are, appear in the bottom box. These scores represent the distance between the user's score and the mean and standard deviation of the scores of a group of male and female native speakers saying the same sentence. Different speakers speak at different rates, so we obtain relative duration by comparing the duration of one vowel to the next, thus the arrow and the parts of words at the heads of the columns. Users receive information about what they have just said (e.g., <-OK, <-SHORT, or <-LONG, indicating judgment on the preceding vowel). Whole syllables are shown as opposed to individual vowels which, by themselves, would be difficult to comprehend. In preliminary tests of the system with 12 foreign graduate students, it took an average of three trials for a student to get all "<-OK"s. These tests consisted of a 20-minute session for each student (there was a set of 10 sentences, all dealing with stressing the auxilliary verb). The system pointed out only incorrect durations, never calling a segment long or short when, according to an expert teacher, it was of normal duration. There were no system crashes and response time was about 1.5 times real time.

In post-trial interviews, students said they had the impression of "being in charge", being able to choose what function to use next and how often to use it. Most of the students were reluctant to stop when they were told that the test was over.

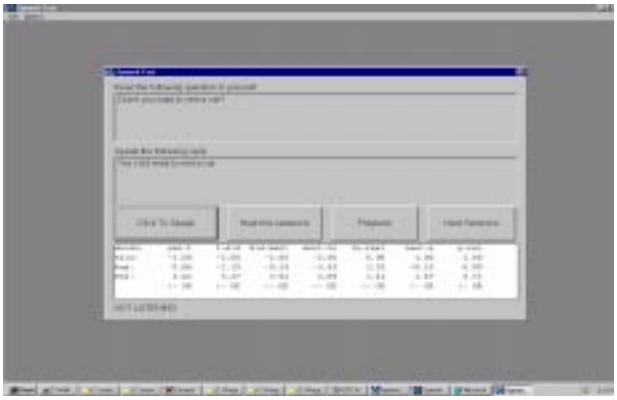

Figure 1. Version 1 Duration correction screen

2.2 Phone correction

Phone correction demands more on many levels. We have concentrated our efforts on two: recognizer implementation and the interface. For the recognizer, we obtain results from forced alignment as in the duration module, are in the process of combining scores from the recognizer as well. Since results are preliminary as of this writing, we will not describe this further here.
For the interface, it is no longer sufficient to point to a syllable and say "short", for example, to correct it. The user implicitely understands how to make a syllable shorter, but does not do the same with a foreign phone. More detailed articulatory instructions are necessary. Several versions of the instructions have been written for use in the different interfaces, each adapting to the user. Details of the versions of the instructions are given in 3.3 below.

3. ADAPTING TO THE USER

3.1 "Golden" voices

Most systems provide the user with one and only one voice to imitate during the exercises. In the duration module's first version, there was only one female speaker. But male speakers with low pitch had difficulty using this speaker as a model. We are starting to offer several male and female speakers as models. The user can choose one he is comfortable with and can change at any time.

3.2 Adaptation to user learning strategy

Self confidence is also enhanced when the system is adapted to the individual user. We are adapting our system to differing learning strategies.

Much past training has been based on showing the student how to articulate new sounds with illustration. It was believed that visual/physical training was necessary to teach the new sounds. Work by [1] proved that new sounds can also be taught by perception alone. Japanese speakers were trained to hear and pronounce the r/l difference in English by only *listening* to instructions and minimal pair examples. This implies that there may be more than one learning strategy; some students may learn better "by ear", others visually (articulatory instructions on the screen). We have been redesigning our interface to provide three corrective feedback options: only aural, only visual, and a combination of aural and visual.

Since many users may not know which strategy suits them best, we have developed an automatic test. It has four parts: 1) there is a set of differently colored buttons with corresponding tones that have fixed places on the screen (see Figure 2). First only one tone/button is played, then two, etc. and the user must imitate that series exactly. When the user repeats the series correctly, a new series, one element longer than the last, is presented. This continues until the user makes an error. The system records the number of elements in the longest correctly repeated series and response times. 2) the same screen without sound. Data is recorded in the same way as before for all games. 3) there are three sounds (door slam, frog, bark) with corresponding aligned buttons that constantly change position on the screen. 4) there are 3 tones with corresponding 1-2-3 buttons which also change position. We postulated that a user who did better, for example, on the second game (much longer "best" series and/or shorter response times) should respond better to visual training and vice versa. A user showing no clear preference for one or another would use a combined method.

Eight users took this test. They were instructed to play each of the four parts several times; we retained the longest correct series for each part. After they played, they were given five questions from an intuitive questionnaire commonly used in second language acquisition classes [2]. For example, *"I understand directions better when a) the teacher tells them to me; b) I read them; c) no preference either way"*.

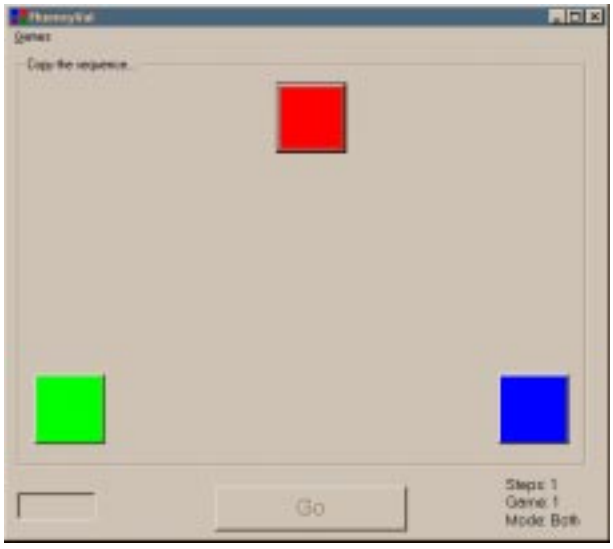

Figure 2. Learning strategy test 1

Figure 3 compares their intuitive responses to test results. Intuitive responses are shown on the solid line. The difference between the aural and visual responses is expressed as a percentage of all answers. A positive number means the user is more aurally inclined; numbers close to zero mean no clear predilection. The long dashed lines (items/soundx - visual) represent test results for the maximal number of items in the longest correct series. They are expressed as:

$$Ix = (Sx - V) / Sx \qquad 1$$

where Sx is the maximal number of sound x items and V is the maximal number of visual items. Again, positive numbers indicate aural inclination, etc. The short dotted lines (duration/soundx - visual) refer to the mean pause time for the "best round" of each. The percentage was calculated as in Equation 1 above, but since longer duration is an indication of a harder task, the sign of Ix was reversed.

The mean response time tends to correlate better than the number of items with the intuitive questionnaire. The use of sounds with linguistic significance (sound1) can't be compared to pure tones (sound2) due to the size of the statistical sample. In general, for mean response time, the tests seem to reinforce each other - if a user did better on one sound game than on the visual game, he also did better on the other sound game.

Due to sparse data, it is hard to surmise why the game and intuition were didn't match for user *plc*. It is possible that the intuitive questionnaire is not "ground truth". Some may not respond well to an intuitive questionnaire (*mhl* had a majority of "neither" responses). In this case, the game results

could be closer to reality. We can test this assertion by determining if *plc* learns better when given aural instruction.

It is also possible that, although the game seems to be a good indication of aural/visual tendency for most users, this may not be true for some portion of the population. This needs to be verified on a much larger user population. Another interpretation is that *plc* may not test well on a task using memory as a variable.

If the test does indeed give promising results when compared to actual learning and with a larger population, then interfaces other than language training ones may also benefit from the use of this test.

3.3 Corrective instructions

For the three learning strategy interfaces, we wrote corresponding sets of instructions to be used for phone correction. Let's take the example of teaching the pronunciation of "th" in English. The user is given a choice of elements such as head cut and hearing minimal pair sentences that appear on the screen during the exercise, as well as the choice of a verbose or less verbose explanation. Examples of possible configurations:

Case1: "visual learner" - accompanying a front view video and a head cut drawing, "Make an "s" sound and then move your tongue forward along the back of your front teeth about midway down".

Case2: "mixed learner" - accompanying a head cut drawing and a button to push to hear minimal pairs ("sin/thin", "sank/thank", "soar/Thor" minimal pairs), "move your tongue forward from the 's' position down the back of your teeth".

Case 3: "aural learner" - minimal pair button alone.

Finally, we should construct a user profile with information such as perferred learning strategy and performance on past exercises, automatically consulting this constantly-updated information to adapt the system to the user.

4. CONCLUSION

We have described user adaptation in the Fluency project, where emphasis is put on augmenting self-confidence as a way to obtain a more effective pronunciation trainer. We especially described the adaptation of the interface to the user's specific learning strategy and hope to include this information and others in a student profile.

5. REFERENCES

[1] Akhane-Yamada, R., Tohkura, Y., Bradlow, A., Pisoni, D. (1996). Does training in speech perception modify speech production?, *Proc. of ICSLP'96*, Sep. 96, Philadelphia.

[2] Brown, H.D., (1991*). Breaking the Language Barrier*, Intercultural Press, Yarmouth, Mass.

[3] Celce Murcia, M., Goodwin, J. (1991). Teaching pronunciation, in Celce Murcia (Ed.), *Teaching English as a Second Language*, Heinle and Heinle.

[4] Eskenazi, M.. (1996). Detection of foreign speakers' pronunciation errors for second language training - preliminary results. *Proc. ICSLP'96, Philadelphia*.

[5] Laroy, C. (1995*). Pronunciation, in Resource Books for Teachers*, Oxford University Press.

[6] Price, P., Rypa, M., (1998). Speech Technology and Language Learning: Some examples from VILTS the Voice Interactive Language Training System, *Proc. AATOLL Conference*, Honolulu HI, Feb. 1998.

[7] Ravishankar, M. (1996*). Efficient Algorithms for Speech Recognition*, Ph.D. Thesis, Carnegie Mellon University, Technical Report CMU-CS-96-143.

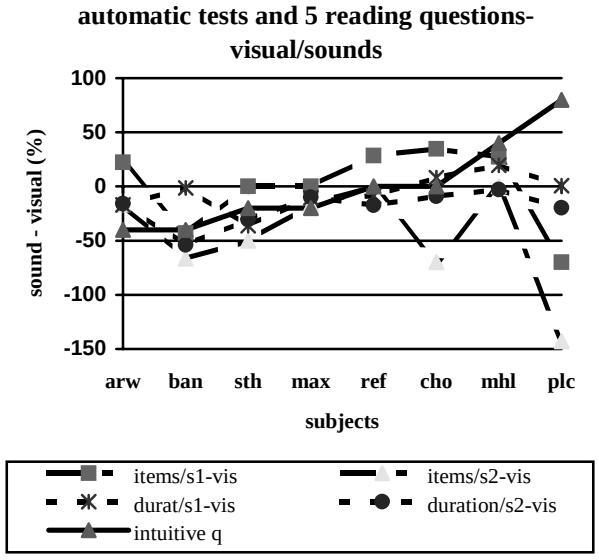

Figure 3. Results of test for 8 users