The Fluency Pronunciation Trainer

Maxine Eskenazi and Scott Hansma

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, Pa. USA
max@cs.cmu.edu, hansma@cs.cmu.edu

Abstract

In this article we describe the basis of the Fluency project for foreign language pronunciation training using automatic speech recognition. We describe the theoretical base, the interactive duration correction module, and our work toward adaptation to the way in which the user learns best. We show results in preliminary tests of the latter, and discuss future directions of the project.

1. Introduction

The goal of the Fluency project is to create a pronunciation trainer for foreign language learning. It uses automatic speech recognition and follows basic principles in foreign language learning research. It will offer correction of both phonetic and prosodic errors and provide user-adapted interfaces. Herein we present duration correction as well as preliminary results in adapting the system to individual learning strategies. While the use of the recognizer for learning grammatical structure, vocabulary and culture is also important, this is not the object of Fluency, exercises involving these language levels will be mentioned only as they are used for pronunciation. The reader is referred to [11] for information about the VILTS project concerning training on these other levels. Pronunciation training is important; even if grammar and vocabulary are completely correct, communication *cannot* take place without correct [5]. Poor phonetics and prosody distract the listener and impede comprehension of the message. Our efforts are therefore concentrated in this area.

Fluency *only* points out errors where it can also provide corrective feedback, showing the user how to correct himself. The feedback, and indeed the interface in general, are designed to make the user feel comfortable with the system and self-confident, a key to continued success in speaking a foreign language [5].

2. Basic Principles of Fluency

2.1 Creating self confidence

Present techniques to boost student confidence [9] consist of correcting only when necessary, reinforcing good pronunciation, and avoiding negative feedback. Avoiding incorrect feedback (for example, telling a student he was wrong when he wasn't) is a major challenge to the use of automatic speech processing; a small margin of error has usually been acceptable in speech applications so far.

Fluency does not pass judgement on the user, or on his way of speaking. Rather, the system pinpoints specific items to be worked on. Although numbers presently appear on the screen for development purposes, scores will not be shown in future versions of the system.

2.2 Active production of speech: elicitation

In past automated language learning, students have had a passive role. That is, they have either had to repeat a sentence they heard, or choose one of three to four written sentences [3] to be read aloud. In both cases, the answers are already constructed (vocabulary chosen and syntax assembled) - students have no practice in constructing their own utterances.

We [6] have developed a solution that enables users to participate more actively. Automatic speech recognition has worked in language tutors so far, due to the fact that the utterances had been known ahead of time (read off the screen), and fed to the recognizer with the speech signal. The system uses forced alignment to match exemplars of the phones it expects (pre-stored in memory) against the incoming signal (what was actually said). It is still possible to predict what students will say to satisfy the needs of the recognizer, while giving them an impression of freely constructing utterances on their own. This can be done by using elicitation techniques, similar to the drills that are the basis of methods such as British Broadcasting Company (BBC) [2] and Audio-Lingual Method (A-LM) [13]. Several studies have been carried out to determine whether specifically targeted speech data can be collected using elicitation [8], [10], [7]. Within a given, carefully constructed exercise, we know that (with cooperative students a very small number of answers are possible after a given elicitation sentence.

The exercise below, from a set of exercises designed for the FLUENCY project, is an example:

Sentence Structure and Prosody Exercise:

System: When did you **meet** her? (**yesterday**)

User: I met her yesterday.

System: 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.

The technique provides fast-moving exercise for the students, making them active, rather than passive speakers. Later, during a real conversation, when they need to build an utterance, they will have acquired the necessary speaking experience and automatic reflexes. They can then speak rapidly, *in pace* with the conversation.

2.3 Providing corrective feedback

Teachers point out incorrect pronunciation at the right times during exercises, infrequently, in order to avoid discouraging the student. They *do* intervene soon enough so errors are not repeated several times, becoming hard-to-break habits. Helpful feedback implies that the correction will give students the tools to deal with other aspects of the same pronunciation problem later on. Pointing to an error without feedback as to how to correct it leads to user trial-and-error and is useless.

2.4 Prosody as well as phonetics

When a student starts to learn a new language some time is usually devoted to learning to pronounce phones that are not present in the native language. Experience shows that a person with *perfect* phone pronunciation who lacks correct timing and pitch is very hard to understand. The "song" that the speaker "sings" while emitting a string of phones is the glue that holds the whole message together, guides the listener along, indicates where important content words are, disambiguates parts of sentences, and enhances the meaning with style and emotion. Ideally, prosody should be taught from the beginning. The aspects of prosody a speaker needs to improve on are its three distinct components: fundamental frequency, duration, and intensity. By exercising one component at a time, specific errors can be explained, practised and understood. Then combinations of components can be worked on until all three elements are made to work together. The two types of pronunciation errors are totally different in nature and 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 also differ in the two languages. In prosody, the elements are the same in each language - speakers know how to vary fundamental frequency, duration and intensity - but the relative importance of each of the three, the meanings linked to each, and the types of variations used differ from language to language. For example, variations of intensity are used much less and show less contrast in French than in Spanish.

In error detection, the methods differ as well. Given the sentence that the speaker was to say, the speech recognizer (in "forced alignment mode") can return the scores of the words and the phones in the utterance. By 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 [3]. For prosody errors, only duration can be obtained from recognizer output. That is, when the recognizer returns the phones and their scores, it can also return the duration of the phones. It is important that measures be expressed in *relative* terms (such as duration of one syllable compared to the next) since speakers vary greatly in individual intensity, speaking rate and fundamental frequency within the constraints of the given language.

3. The Fluency system

In past work [6], we have shown that it is 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 that we have designed takes advantage of our findings on duration. We have chosen to work on duration first for three reasons: 1) our belief that prosody is as important as phonetics; 2) the corrective feedback for duration errors necessitates less elaborate implementation since we do not have to show articulator placement, for example; 3) duration has the lowest associated error rate, enabling us to start with a system that functions as dependably as possible, to establish student confidence and serve as a building block for future work.

3.1 The duration trainer

We correct duration separately form the other elements of prosody since it is our belief that the user will not be able to correct himself (will not understand what needs to be corrected) if he does not have precise elements and precise instructions as to how to correct them. When the duration patterns are learned, then the user can proceed to pitch, etc., and then, when all elements work well individually, the user can be given combinations of them to try. This progressive attitude should ensure less error and thus faster progress.

We use CMU's SPHINX II automatic speech recogniser [12] in forced alignment mode to furnish duration information.

In this module, the speaker is asked to say a sentence. The sentence is *elicited* from the speaker by the system. At present, in our proof-of-concept system, there is a base sentence 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 box marked, "Speak the following reply". These two boxes will be replaced by a "talking head" in the next version of the interface.

The user clicks on the "Click to speak" button and says his sentence, clicking on the button again when he has finished. The "closed mike" can eventually be replaced by an "open" one, but again, we chose whatever contributed to the lowest overall error rate (and so we avoided marginal error associated with silence detection).

After saying his sentence, numbers (which, again, will disappear in the next version of the system) and arrows to errors appear in the lowest box. If the voiced segment is correct, an "OK" appears. If it is not, an arrow and "LONG" or "SHORT" appears. In order to deal with the fact that different speakers speak at different rates, we compare the duration of one vowel to the next, the arrow points to one of the two syllables above the column. Although only vowel duration is measured, we show the whole syllable for easier comprehension

Upon seeing how well he did, the user can listen to what he said (the "Playback" button) and/or listen to a native speaker (the "Hear this sentence" button). He then can go back and forth as often as he wishes among the three options until he is satisfied with his results. Preliminary tests of 12 foreign students show that an average of 3 recordings were all that was necessary for the student to obtain results that enabled him to get all "OK"s in the bottom box.

In the preliminary test, a 20-minute session for each of the 12 students, the system pointed out only incorrect durations. It never called a segment long or short when, according to expert teacher judgement, it was of normal duration. There were no system crashes and response time was about 1.5 times real time.

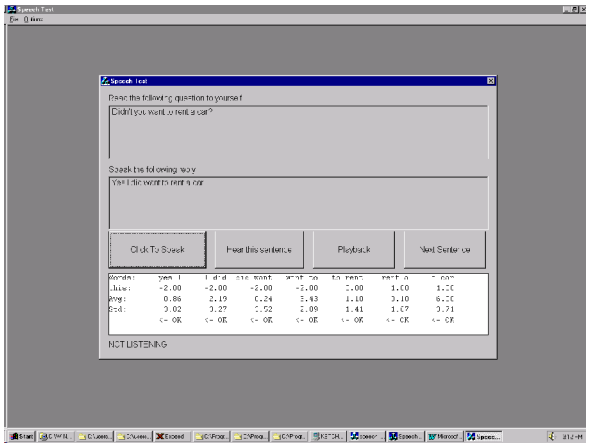

Figure 1. Fluency duration trainer main screen

3.2 Adaptation to user learning strategy

Confidence is also enhanced when the system is adapted as much as possible to the individual user. Amongst other efforts in this direction, we have begun to adapt our system to differing learning strategies.

Much past training has been based on showing the student how to articulate new sounds with instruction and illustration. It has been believed that visual/physical training was necessary to teach new sounds. Recent work by [1] has proven 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 simply listening to instructions and to carefully chosen minimal pair examples. This implies that there may be more than one learning strategy; some students may learn better "by ear", others with "visual correction" (articulatory instructions on the screen). We are redesigning our interface to provide three corrective feedback options: only aural, only visual, and a combination of aural and visual. Since many users (especially young ones) may not know which strategy suits them best, we have developed a game. It has four parts: 1) there is a set of differently colored buttons with corresponding tones that have fixed places on the screen. 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. The series gets longer and longer until the user makes an error. The system records the number of elements in the longest correctly repeated series and response times. 2) there are colored buttons, but no sound. Data is recorded in the same way as before for all games. 3) there are three sounds (door slam, frog, bark) with corresponding buttons that constantly change position on the screen. 4) there are 3 tones with corresponding 1-2-3 buttons which also change position. We postulate that a user who does better, for example, on the third and/or fourth games (much longer "best" series and/or shorter response times) should respond better to aural training and vice versa. A user showing no clear preference for one or another would be offered a combined method.

We tested this game for 8 users. They were all instructed to play each of the four parts several times and we retained the longest correct series for each part as the "best round". After they played, they were given five questions from an intuitive questionnaire [4] commonly used in second language acquisition classes. 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 compares their intuitive responses to what they did in the game. The intuitive responses are shown on the solid line. The difference between the aural and the visual responses is expressed as a percentage of all answers. Therefore a positive number means the user is more inclined to aural learning; numbers close to zero mean there is no clear predilection. The long dashed lines (items/soundx - visual) represent game 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, a positive number indicates 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.

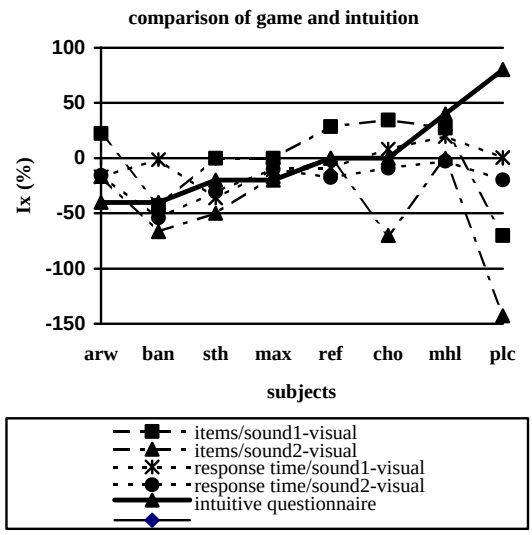

Figure 2. Test of the learning strategy game

3.2.1 Discussion of results

There are several possible interpretations of the above results. For most of the subjects, the mean response time correlates better with the intuitive questionnaire results than the longest correct series (items). We cannot compare the use of sounds with a linguistic significance (sound1) 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 so ill-matched for user *plc*. Several interpretations are possible. First, it is possible that the intuitive questionnaire is not "ground truth", even for this group of students. 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 only test the verity of this assertion when we have created the three different interfaces and we determine whether *plc* indeed 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, it may not be the case for some portion of the population. Again, this hypothesis can only be verified on a much larger population of users. Another interpretation is that *plc* does not test well on a task where memory is a variable.

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

4. Conclusions

We have presented the theoretical underpinnings of the Fluency project, the duration training proof-of-concept interface, and our first steps to adapt our interfaces to individual users. First results are extremely encouraging; the duration trainer has been dependable and well-accepted by students. Future work lies in ameliorating the duration trainer interface, creating a phone trainer, and refining our learning strategy test and incorporating these strategies into the trainers.

Acknowledgements

We would like to thank Monique Semp and Randy Warner for their help on the interface and John Corwin for his contribution to the duration correction module.

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] Allen, W.S. (1968). *Walter and Connie, parts 1 - 3*. British Broadcasting Corporation.

[3] Bernstein, J., Franco, H. (1995). Speech recognition by computer, in N. Lass (Ed.), *Principles of Experimental Phonetics*, Mosby, 408-434..

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

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

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

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

[8] Isard, A., Eskenazi, M. (1991) Characterising the change from casual to careful style in spontaneous speech. *Journal of the Acoustical Society of America*. 89:4:2.

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

[10] Pean, V., Williams, S., Eskenazi, M. (1993). The design and recording of ICY, a corpus for the study of intraspeaker variability and the characterisation of speaking styles. *Proc. Eurospeech '93*, Berlin,. 627 - 630.

[11] 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.

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

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

[2] Lindgren A (1995). Title of Paper, *Important Journal*, 43: 123-128.

[3] Lindgren A (1995). Title of Another Paper, *Less Important Journal*, 16/3: 123-128.

[4] Strindberg A (1998). *Important Book*, University Press, Paris, 1998.