

The Fluency Pronunciation Trainer: Update and user issues

Maxine Eskenazi, Yan Ke, Jordi Albornoz, Katharina Probst

Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, Pa.15213, USA
max@cs.cmu.edu, yke@andrew.cmu.edu, jordi@andrew.cmu.edu, kathrin@cs.cmu.edu

Abstract

The article describes the present state of the Fluency pronunciation trainer and deals with issues concerning the role of the user that have come up when choices had to be made about the system's interface. We first refer to several CALL systems in order to examine user and system initiative in controlling the course of events in a session. In light of this, we then present new additions to Fluency: choosing speakers, using a rejection threshold, limiting error correction, giving listening information, and finally, creating an authoring kit.

1. Introduction

The Fluency Pronunciation Trainer [1], [2] detects foreign language pronunciation errors and offers correction hints. Work in the past has mainly focused on duration and phone error detection. More recently, the interface's feedback has been expanded, giving better help in correcting errors. An authoring kit also allows teachers to easily add new exercises.

This paper deals with interaction issues that we addressed as we expanded the Fluency system. At each step there have been choices in the way information is presented to the user. Some of these choices involve pedagogical principles for which there are not, to our knowledge, any empirical results on actual CALL systems to guide our decisions. User tests were carried out for some of these aspects. For others, we chose the form of implementation which would allow the user to decide what to do, or at least give her the impression of being "in charge". This is intended to heighten self-confidence, a strong factor in language learning success [3]. We will discuss the implication of that decision as we describe recent additions to the Fluency system. We will begin by discussing choices made by authors of other CALL systems presented at the last STiLL workshop as a means of putting the role of the user into perspective.

2. The listen/speak sequence

CALL systems that are used to improve pronunciation usually call upon the user to imitate speech she listens to from the system. Sometimes the user responds to a teacher prompt or question. For each type of interaction, system designers must decide whether the user or the system has the upper hand. Does the system tell the user when to speak and then automatically decide when to

send a speech signal back? Or does the user initiate everything? Or is there a mixed-initiative solution?

In designing Fluency to let the user decide what to do next, we have found he does not always know best (see an example later in this paper). Looking at CALL literature, we can expect better quality interaction with a mixed-initiative system where the user has control of some events and has the *impression* of controlling others.

In general, we can distinguish three types of events that can be system- or user-initiated or both: 1) the sequence in which events take place, 2) the timing of the events, and 3) the repetition of an event.

2.1 Sequence of events

In the first case, the order in which the events take place can be fixed or flexible, according to what the user said, or to what the user wants to do next. If the sequence is to resemble a conversation, for example, the order is fixed: a series of system-user interchanges where each side produces one and only one utterance (an utterance may be a set of several sentences) in a system-directed order. But in the case where pronunciation is being practised, the order may not be as crucial.

In a mixed-initiative situation, the user can be in charge of choosing a starting point (what to work on) then the system may take over. In the case of total system initiative, the system may even decide on the starting point, based on a running record of past user performance. Faced with these alternatives, the final choice for a given system can also take the nationality of the end users into account, as certain forms of interaction are more desirable than others in some cultures. For example, the user who is learning a language that is very different from his own (writing system, morph- vs. word-based, etc., giving many more items to attend to) may welcome the system-initiated sequence of events, giving him the chance to focus on other things.

2.2 Timing of events

The time between events can be variable. It may depend on either the user or the system. If system designers decide that the user must answer within a limited amount of time, for example, in order to create a tempo for the exercise that paces the user's interaction, then it would be system-initiated. Such a system must deal with the fact that each individual requires a different amount of time to integrate information and to prepare to say something. This is a point where it is best to let the user decide, at least for her own speech, when to

start. The timing of the *system's* prompts can be user-initiated or can come at a reasonable time after a correct user utterance, depending on the goals for the system.

2.3 Repetition of events

The user may want an event, such as a correct user utterance, to be repeated. The user may also want to hear the system prompt again. In the case of a dialogue, the system will move the user on to the next interchange in order to keep up a communicative chain and a realistic pace of conversational speech, giving no opportunity to hear or practice a token again. But in many cases of pronunciation training, the user may wish to try something over several times. This desire to reinforce a newly-learned item differs from individual to individual. In order to be as effective as possible, a system should be adaptable to different ways of learning and allow for enough repetition for a user to become comfortable with the newly-learned item. The automatic system can offer advantages over classroom teaching here, where time for the teacher to listen to individual utterances is split amongst the students and necessarily limited.

In order to illustrate the above three points, we give examples below from the last STiLL workshop, they come from papers that described the sequence of system events, illustrating the variety of user- and system-controlled events. There are two types of systems that can be distinguished: *pronunciation* systems and *conversation* systems. Although we are concerned with the former, we have included a few examples of the latter for comparison in light of what the system is teaching.

Three systems exercise the student's *conversational* skills. [4] has the student give a response to a system prompt. This response automatically triggers a new system prompt with no user freedom in choosing the sequence or repetition of events. This is justified since the system gives an impression of participating in a conversation and is setting the pace of the dialogue. [5] describes two conversational systems. In the Airport system, the student speaks then, depending on what is recognised, the system decides what will come in the dialog. In the Hike system, the student has the possibility of controlling the repetition of events by asking the system to repeat as often as she wishes. The choice of what to do is limited, so this ability to stop and listen over and over gives the user some feeling of being in control.

Two other systems teach *pronunciation*. In [6], the student starts a *listen/speak* sequence by clicking on a syllable that the system pronounces. It then listens for the student's speech. The student therefore has only partial control of the sequence of events, and practically no initiative as to when to speak. If there is an error, the system initiates an information-giving/speaking sequence, thus controlling the repetitions. Since the student has chosen what to work on, he has the

impression of making choices even though the system controls the sequence of events.

In [7], the student clicks to initiate recognition. The choice of what to work on, sentences or phrases, is also up to the user. The system gives feedback and passively waits for the next user decision as to what to do. The user is very much in charge here.

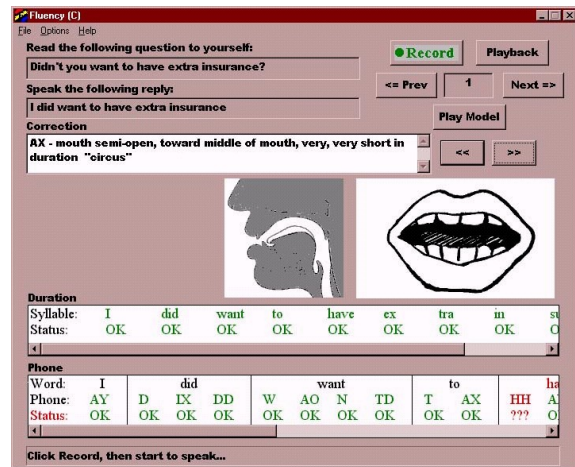
In the Fluency system, the user can either listen to a native speaker, or speak to the system. After seeing error results, the user can choose from several correction options and listen to her own speech or listen to a native speaker. The user has been given the task of deciding how long to remain on one item and what sequence of events corresponds to the way he learns best.

Table 1 compares user/system control for the six systems mentioned above.

Table 1. Control of events in six systems (respectively, Del=[6]; DavA=[5] Airport; DavB=[5] Hike; Mea=[4]; Neu=[7]; Flu=[2]) (s=system;u=user;m=mixed initiative)

	Del	DavA	DavH	Mea	Neu	Flu
seq.	s	s	m	s	u	u
timing	m	s	u	m	u	u
repet.	s	s	u	s	u	u

Figure 1. Fluency main screen.



3. Changes in the Fluency system

This section will present: the choice of speakers, the use of a rejection threshold, the choice of error correction, new listening information, and Fauthor, the authoring kit with references to system-user interaction.

The Fluency system shows a sentence on the screen for the user to pronounce. The user can say it; and then, after seeing where she made errors in duration or phones, listen to a native speaker, listen to herself, or try again. In our experience, users choose a variety of combinations of these possibilities, some also decide to work over and over on one utterance for up to several

minutes. There is advice in written form as to how to place articulators to make a sound and sidecuts and front views of a head as an illustration. Instructions include simply how to form the sound as well as how to start at a sound in L1 and progress to a sound in L2.

Figure 1 is a screenshot of the Fluency main screen.

3.1 Choosing speakers

Fluency users listen to native speakers (“golden speakers”) for comparison with their own speech and for imitation. The system now offers a choice among several golden speakers. We recently [8] examined whether the choice of which golden speaker to imitate made a difference in learning. Users were divided into three groups: those who could choose their golden speakers, those who were matched to their golden speakers, and those who were mismatched to their golden speakers. Users who were matched to their golden speakers performed better on a duration learning exercise than the other two groups of users. We noted that users who chose their golden speakers often chose speakers who were very different from themselves (female users chose male speakers, people with low average F0 chose golden speakers with much higher F0). This led us to conclude that the decision of which golden speaker a user should imitate should probably be left to the system (obtaining a sample of the user’s speech in his native language and comparing it to the several golden speakers to automatically determine the closest match).

3.2 A rejection threshold

When a user is asked to respond to the system, we assume he will co-operate and say what is asked for. Even so, several unexpected events may take place.

First, the user can hesitate (“That’s a_a_a table.”) or restart (“That’s a that’s a table.”). In these cases, we believe that the user will perform as he would in a natural conversation and repeat what he intended to say. In this case, the user simply clicks on “record” again and corrects himself. We believe that this is one task that is best left in the hands of the user than addressed by implementing heavy automatic schemes that would add error to the system responses.

The user can also make a mistake, saying something different from what is expected. As above, we could expect self-correction, but in this case, the user may not be aware that she is mistaken. It would be unnatural for the system to offer error detection since the errors would not correspond to what the user was to say - the interface should detect that the expected utterance was not said and ask the user to say the sentence again. (Note that Fluency always expects the user to say a predetermined sentence.) We have implemented a rejection threshold that combines forced alignment and phone recognition results in order to determine the likelihood that the expected utterance was pronounced.

3.3 Choosing which errors to correct

There can be many pronunciation errors in an utterance. It is annoying to be corrected on a lot of errors in one utterance. There must be a selection of what to correct in order not to interrupt the flow of the lesson. In order to limit the number of errors corrected in a given utterance, we have implemented two types of filters.

The first filter lets the user decide how many errors he wants to have corrected in any sentence. He can choose the worst error, the two worst errors, the three worst, etc. Worst error is defined as the phone or syllable error that has the worst score of all of the phones or, respectively, syllables in the utterance. The system will then show this number of errors or less if it did not detect that many.

The other filter follows the principle of teaching just one thing at a time. For example, if the exercise that is being worked on is about pronouncing the “th” sound as in “thin”, this filter will only let the interface show errors on the phone “th”, and not, for example, on the “n” at the end of the word. The student can always go back and work on other sounds later.

We have left access to these two filters open to the student so far, but we plan to determine the second filter’s contents automatically according to the exercise in use. During authoring of an exercise, the filter will be chosen and stored along with the exercise information.

3.4 Listening information

While Fluency has so far been oriented toward the use of images and instructions as to where to place articulators, recent work [9] has shown the value of listening to minimal pairs or enhanced minimal pairs. We have not yet implemented enhanced minimal pairs, but we do let the user hear one single word in a sentence by clicking on it, or, again by clicking on a word, hear that word in a minimal pair (for example, in an exercise on “th”, if the user clicks on “thin” and has asked for a minimal pair, she will hear her golden speaker say “sin, thin, sin, thin”). The user, not the system, triggers this at present. We are examining logfiles of user sessions to determine whether this feature is actually used. If it isn’t, we may have the system initiate playback of the minimal pairs at some point in the interaction.

The use of minimal pairs brings up two interesting pedagogical questions. First, should the user listen to minimal pairs made up of a target L2 sound preceded by a sound in his L1 (such as “s” in the “sin/thin” example), or should he hear enhanced sounds only? We are designing tests on Fluency to answer this question.

The second question concerns the adaptation of the system to the user. In learning the “th” sound, for example, speakers of Japanese will pronounce the word “birthday” as “birsday” before learning the sound whereas speakers of Serbian will say “birtday”, and speakers of Russian and of Afro-American English

would say “birfday”. This implies that we need to know the native language of the user and adapt the minimal pair (and instructions on how to go from a sound in L1 to a sound in L2) to the characteristics of the relation between that language and the target language.

3.5 Authoring new content

It is important for anyone using Fluency, either for learning or to test a pedagogical theory, to be able to easily introduce new content into the system. Without a separate authoring interface, the person would have to know how to program and would need to understand how the whole Fluency system works. Fauthor is a kit that does not require this knowledge and thus makes most users capable of authoring. We should note that, at this writing, Fauthor presumes that the user knows the phonetics of the language.

The user is asked to give a name and short content description for each exercise. Then the user proceeds to type in the utterances in the exercise. The system checks the recognition dictionary to make sure that all words are present and, in the case of an out-of-vocabulary word, guides the user in making the dictionary entry. When all of the utterances are typed in, Fauthor guides the user in recording speakers for the exercise. At the end of the recording, Fauthor automatically calculates the statistical representation of the speech and the exercise is entered in the Fluency system. It also chooses the golden speaker sentences. Thus, when the user has finished recording and clicks on a button, Fluency is ready to use for the new exercise. The user has not needed to open or move files, or carry out any other actions that would not be known by the average language teacher. The minimal pair hearing option should be fully automated in the near future.

4. Discussion

We have shown how a variety of system designers have chosen to design user interaction. We have also shown the changes to Fluency and discussed them in light of the role of the user. There is no simple answer concerning the role to assign the user. It varies according to the system goal (dialogue, pronunciation, etc.) and the measuring stick that the system creators choose. User confidence is one of several criteria that may enter into this decision. Here are samples:

- Does the choice follow the pedagogical principles put forward by X? (The goal here being to create a system that teaches language according to a targeted set of pedagogical principles.)
- Does the choice afford the greatest sense of user self-confidence? (The goal being to afford progress by affecting the psychological state of the user.)
- Does the choice give the best overall scores in improvement (learning)? (The goal being to choose the option that affords the best results in a set of user tests.)

- Does the choice afford more continued use/acceptance of the software (interacting with it more often and over a longer period of time)? (The goal here being the final usability of a product, independently of its other characteristics.)

These criteria imply trade-offs, such as wanting more precision or more user control. We have seen that the user does not always make the optimal decision, so more user freedom implies some loss of performance improvement.

5. Conclusion

In the future, we hope to make Fluency even more flexible and test a variety of pedagogical tools and theories with it. We will also give the user less direct control of several aspects (such as the choice of the golden speaker), while creating the impression of control.

References

- [1] Eskenazi, M (1996). Detection of foreign speakers' pronunciation errors for second language training – preliminary results, *Proc. International Conference on Spoken Language Processing '96*, September 1996, Philadelphia.
- [2] Eskenazi, M. and Hansma, S. (1998). The Fluency pronunciation trainer, *Proc. STiLL Workshop*, May 1998, Marholmen, 77-80.
- [3] Celce Murcia, M. and Goodwin, J. (1991). Teaching Pronunciation, Celce Murcia, ed., *Teaching English as a Second Language*, Heinle and Heinle, 1991.
- [4] Meador, J, Ehsani, F, Egan, K, Stokowski, S (1998). An interactive dialog system for learning Japanese, *Proc. STiLL Workshop*. May 1998, Marholmen, 65-68.
- [5] Davies, S. Poesio, M (1998). A CSLUrp-based spoken dialogue system for teaching English as a foreign language, *Proc. STiLL Workshop*, May 1998, Marholmen, 183-186.
- [6] Delmonte, R (1998). Prosodic modelling for automatic language tutors, *Proc. STiLL Workshop*, May 1998, Marholmen, 57-60.
- [7] Neumeyer, L., Franco, H., Abrash, V., Julia, L., Ronen, O., Bratt, H., Bing, J., Digalakis, V., Rypa, M. (1998). WebgraderTM: a multilingual pronunciation practice tool, *Proc. STiLL Workshop*, May 1998, Marholmen, 61-64.
- [8] Probst, K, Ke, Y., Eskenazi, M (2000) Enhancing foreign language tutors – In search of the Golden Speaker, *accepted for publication in Speech Communication*.
- [9] Pruitt, J., Kawahara, H., Akahane-Yamada, R., Kubo, R (1998). Methods of enhancing stimuli for perceptual training: exaggerated articulation, context truncation, and “straight” re-synthesis,

Proc. STiLL Workshop, May 1998, Marholmen,
107-110.

This document was created with Win2PDF available at <http://www.daneprairie.com>.
The unregistered version of Win2PDF is for evaluation or non-commercial use only.