An Empirical Study of the Effectiveness of Speech-Recognition-Based Pronunciation Training

Laura Mayfield Tomokiyo, Le Wang, and Maxine Eskenazi

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

Abstract

We have tested a fixed version of the Fluency foreign language pronunciation trainer to determine whether users actually improve their pronunciation through its use. The Fluency system uses the CMU SPHINX II speech recognizer to pinpoint pronunciation errors and then gives suggestions as to how to correct them. Users in the experiment we will describe in this paper were to learn to pronounce the voiced and voiceless interdental fricatives /θ / and /ð/, as in "thin" and "that." The choice of these two phones allowed us to attract users from several native languages. In this paper, we present the results of this study, fully describing the test situation and drawing conclusions about Fluency as a testbed.

1. Introduction

In this paper we describe an experiment that was carried out to determine whether using the Fluency system [1] really helps non-native speakers of English improve their pronunciation skills. Fluency is a system that uses the SPHINX II [2] recognizer to detect pronunciation errors and then offers hints on how to correct them. We taught two sounds that are a problem for speakers of many languages when learning English: the unvoiced /θ/ as in "thin" and the voiced /ð/ as in "that". The interface design included user-driven exercises and feedback without scores. Fluency offers both duration and phone correction - in this experiment only phone correction was used. To see if the user's pronunciation could be modified in a relatively short amount of time, each user was given a total of about two hours exposure over a period of two to three weeks.

The goal was not to compare Fluency to any other system, nor was it to validate one pedagogical principle as opposed to another, although these are planned future activities. The conclusions we will draw here concern this system in the configuration described below and only draw a comparison to a classroom situation that was used as the control.

2. Background

Most work in evaluation of pronunciation tutoring systems has focused on establishing whether ratings given by the automatic system match those of a human rater, and if inter- and intra-rater consistency is high. Various rating scales have been proposed. Bernstein [3] used a seven-point scale to rate pronunciation at the sentence level. Neumeyer [4] also rated speakers at the sentence level, but uses a five-point scale. Witt and Young [5] used a binary scale, giving scores at the phoneme level. These schemes successfully showed the correlation between human and machine judgements of pronunciation quality, and that an automatically-derived score can be used for pronunciation evaluation. Our objective is to determine whether pronunciation improves with use of the Fluency system. Ratings given by a human listener (the rating scheme we selected is described in section 3.6) are not compared with ratings given by the system; rather, we look at ratings given before and after training. In this respect, our evaluation is similar to non-ASR based CALL system evaluations. MacWhinney [6] gives several principles to be considered when evaluating foreign language tutoring systems, including appropriate control group comparisons, the process-neutrality of the post-test, the effect of being singled out for special treatment, and the quality of the process (system bugs, e.g.). Chapelle [7] identifies two critical questions to ask when evaluating CALL activities: the kind of language the learner engages in, and how good that language experience is for L2 (target language) learning. Dunkel [8] emphasises the importance of considering student ability, native language group, sample size, and the L2 skill being targeted in drawing any conclusions about CALL system effectiveness.

3. Experimental setup

Choosing the target sounds, we reasoned that if we concentrated on a sound in English that was not present in only one or two languages, we would only be able to draw conclusions about speakers of specific L1s (native languages). We therefore sought a sound that is not found in many different languages. The voiced and unvoiced /θ/ and /ð/, as in "thin" and "that" are not present in such diverse languages as German, Chinese, French, and Japanese. This gave us users from a variety of L1s and, hopefully, L1-independent conclusions.

3.1 The Fluency interface

The Fluency system not only detects errors, but then points the user to where they occurred and offers hints on how to correct pronunciation. In this experiment, those hints take the following forms:

1) hear oneself; 2) hear a native speaker; 3) read how to pronounce a sound; 4) see side headcut and front view of lips; 5) hear the word in isolated form; 6) hear the word as a part of a minimal pair…("sin"/"thin").

The user therefore has the freedom to choose both aural and visual hints at will. Following a user's request, we also provided a mirror to aid in microphone placement and feedback on user articulator placement

The articulation instructions sometimes presumed an L1 starting point of /s/ or /z/. This is not the optimal presentation; while many speakers of Japanese replace a /θ/ with an /s/, speakers of German often replace it with /t/ and speakers of Russian with /f/. In the future, knowledge of the L1 of the speaker will be used to better fit the correction information to the speaker.

3.2 Exercise and test materials

Fluency attempts to help the speaker produce sounds appropriately in context, never in isolation. The exercises and pre/post-test were structured to provide exposure to a variety of contexts. Specifically, we addressed:

1) placement (beginning/middle/end of word); 2) context (only vowels or with consonants) (in word alone or in a sentence); 3) anticipation (was there also a close sound, such as “s” within the word after the target sound - “fifth” is easier to say than “fifths”).

We created 23 exercises to cover these contexts, each having from 3 to 13 items. For each exercise we also created a corresponding test exercise with words of similar difficulty, but containing different words so that users were trained and tested on disjoint sets of words. There were 143 occurrences of /ð/ and /θ/ in the test exercises and 157 occurrences in the training ones. Subjects in both the Fluency group and the class group used these exercises exclusively.

Duration of exposure to the exercises was chosen to allow: a sufficient amount of time for learning to take place, be reasoned about and assimilated in everyday experience; and accordance with subjects’ varied schedules, short enough to ensure everyone could come every time. Preliminary trials of the exercises showed that an average user could finish all 23 exercises in about 140 minutes. The users in both groups were carefully timed and all given a total exposure of 140 minutes (with a 3-minute leeway). Subjects had six sessions, each running 20-25 minutes for either two (3 sessions per week) or 3 (2 sessions per week) weeks.

3.3 The classroom control group

Designing a lesson plan for the control group was challenging. As this was a controlled experiment, we needed to fix as many variables as possible, varying only the source of the guidance in completing the task. However, it can be argued that comparing online-tutor instruction to a scenario that would not ordinarily happen in the classroom is not very meaningful. We therefore decided to fix the content (the list of words to which the students were exposed) and the task (pronouncing those words aloud), allowing the instructor some latitude to provide a natural, effective classroom experience for the students. The six classroom sessions (averaging 23 minutes) all followed the same format. For the first one to two minutes, the students “warmed up" by practicing articulation of isolated /θ,ð/, with a different pronunciation hint given each session based on the type of phonemic context to be covered. For example, at the third session, for words having /θ,ð/ in consonant clusters, the instructor described how coarticulation is used to make phoneme transitions easier.

The lists of words to be covered that day were then displayed on an overhead projector. The class went through them one by one, going around the room so that each student pronounced a word in turn. For difficult words, each student pronounced the word, giving them the benefit of practice and of hearing other students' pronunciation and the instructor's feedback. There was no time to have each student pronounce every word, so in most cases each student read a new word. After a pronunciation, the instructor either moved on to the next student or asked the current speaker to say the word again. If the incorrectly pronounced word was part of a sentence, the instructor indicated which word was incorrect, had the student try again, and then asked the student to repeat the whole sentence. In general, students were not asked to repeat any given word or sentence more than three times. The instructor's decision to flag a pronunciation as incorrect was based on how the target segment sounded (students were never corrected on sounds other than /θ,ð/), the skill level of the student (requiring more precision for more skilled speakers), and the response of the student to correction (an effort was made not to embarrass the student to the point that instruction was not effective, although giving less shy students more individual attention was also avoided).

3.4 The Fluency group

Fluency group had tally sheets and were told to progress through the exercises from number 1 to number 23 sequentially (increasing difficulty). They wrote down the exact start and end time of each session and which exercises they did. Before the last session, times were tallied and they were told how much time to spend on task in order to make their total approximately 140 minutes. The first session included a short explanation of the functions of the interface with users trying each function on their own.

3.5 Subjects

Most subjects were graduate students. They took the pretest and gave demographic information about length of stay in the US, time they had studied English, etc. All speakers were evaluated for general English proficiency using the SPEAK system [9]. In this system, speakers are given scores from 0 to 3 on pronunciation, grammar, fluency, and overall comprehensibility for a variety of tasks according to strictly defined principles. A score of 3 does not reflect native speech, but rather speech that is completely comprehensible yet detectably non-native.

Subjects were divided into two groups, the class group and the Fluency group. Students were paired (one from class, one from Fluency) by: 1) L1 (into the four categories defined by the Foreign Service Institute) and 2) level on the pre-test. The FSI categories represent “distance” from English - with, for example, French and Spanish in Class 1 and Korean and Japanese in Class 4.

This enables us to match students without matching exact L1's due to the low number of volunteers from some countries.

There were 16 subjects (8 in each group). This low number of subjects has us presently recruiting others so that we may soon report on larger numbers. We have actually trained and tested 34 subjects, but only the ones we report on here could be paired by L1 *and* pretest scores. The following chart shows the characteristics of the participants, grouped by pairs.

Usr	L	L1	M/F	GP	pre
SMB	Ger	1	F	F	2.42
EZC	Fr	1	M	C	2.45
TZG	Turk	3	M	F	1.19
AZC	Thai	3	F	C	1.18
ABS	Serb	3	F	F	2.66
MMG	Rus	3	F	C	2.67
HSK	Kor	4	M	F	1.25
ZZW	Chi	4	M	C	1.17
ABZ	Chi	4	F	F	2.23
YZZ	Chi	4	F	C	2.25
CZH	Chi	4	F	F	2.43
YZS	Jap	4	F	C	2.42
RZL	Chi	4	F	F	2.71
SZM	Kor	4	F	C	2.76

Table 1. Users: L= native language L1= Foreign Service Institute group; M/F= male/female; GP= Fluency or class; pre= pretest score

3.6 Grading

We chose to grade subjects on a three-point scale where 1 meant another sound had been pronounced, 2 meant that the sound was not identifiably another phoneme, was still not quite right and would have been corrected by the rater in a class situation, and 3 meant that the target sound had been pronounced. There was one grader, a professional ESL instructor at CMU. She had little knowledge of the project goals and did not know who she was grading or whether it was a pretest or a post-test. We created baselines for the grading in two ways. First, we determined the score of a native speaker by including a pretest by a native female speaker in the data without informing the grader. The native speaker attained an overall score of 2.769. This shows the fairness of the grader and represents articulatory imprecision present even in careful native speech. Second, to determine whether there was a learning effect just by seeing the words in the pre/post-test twice, we had a good non-native speaker take the pre/post-test twice in a row. He scored 2.853 on the pretest and 2.951 on the post-test, giving an overall change of 3.43% and an overall error reduction of 66.67% (p=0.0331[1]).

1 As in all tests in this experiment, we used an ANOVA analysis.

3.7 Experimental Results

Figure 1 plots pretest and post-test scores. Users in the Fluency group are on the left, the class group is on the right. They are plotted in order of ascending pretest scores.

To plot individual improvement more precisely, in view of the maximum amount of improvement possible (error reduction), we used the following statistic:

$$ER_u = (max - post_u) / (max - pre_u)$$

where max is the maximum score of 3, and pre_u and $post_u$ are the pretest and post-test scores for the given subject.

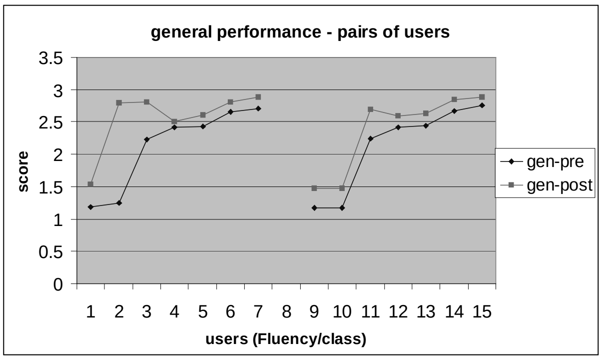

Figure 1. Pretest and post-test scores for all paired subjects

Figure 2 shows error reduction for the paired subjects.

The mean error reduction for the Fluency group is 47.18% and for the class group is 37.52% with although this difference on such a small population is not significant (p = 0.456). There is a lot of variance in the error reduction for the Fluency group (776.333) whereas the variance for the class group is much less (323.308).

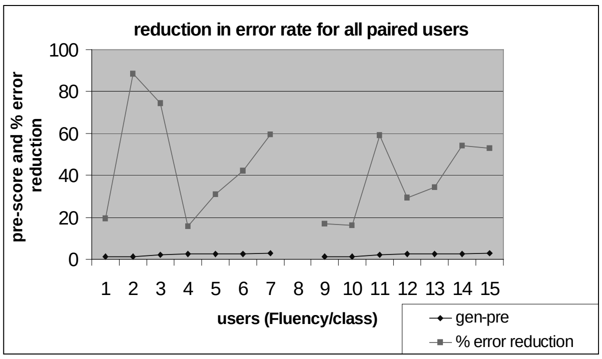

Figure 2. Error reduction plotted with pretest scores for all paired subjects

We also looked at whether the learning effect was the same over the different types of contexts in which the /θ/ was presented. Thus, for middle-of-word contexts, Figure 3 shows the pre and post-test differences. This graph is typical of the results of specific contexts, with some users learning better for certain contexts than others.

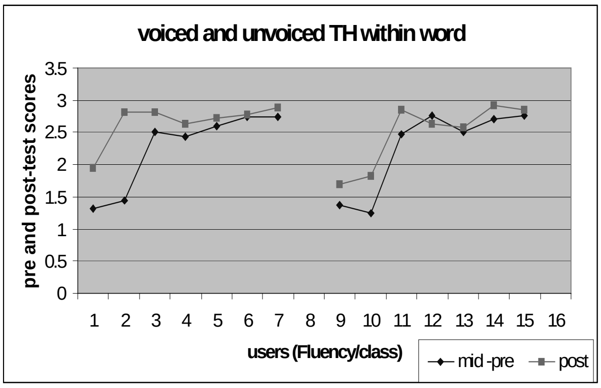

Figure 3. Scores for all paired users in within-word context

4. Discussion of results

The figures from the paired users show that both in class and with Fluency, users can achieve notable error reduction in their pronunciation of /θ/ and /ð/. This result, on speakers who are all well past adolescence, shows that research on pronunciation learning is promising, defeating the commonly held belief that new sounds cannot be learned after the onset of adolescence. We believe that attitude and self-confidence play a large role here.

The variance of the results for the Fluency group is much larger than that of the class group. This may be due to the fact that the human teacher can see if someone is having problems and intervene to solve them. It may also be due to reactions to using a computer - users may be either wildly enthusiastic or fairly negative. A few potential subjects had expressed the wish to be in the class group when they came for their pretests. We tried to get an even number of people who had said this into the two groups. We did not notice any effect of this in the final results.

We looked at the correlation of results with level of pretest. While we had expected that people with a lower beginning level would show more progress, that does not seem to be the case. We also looked at error reduction as a function of native language, SPEAK score and previous exposure to English and found no correlation there.

We examined the contexts and their relation to the results. As mentioned above, in the example of the mid-word /θ,ð/, we found no contexts for which either Fluency or the class, or both, helped much more or much less than other contexts (the mean error reduction varied from about 40% to 53%). Therefore, the system and the exercises we developed seem to show an overall consistency in quality of pedagogical information.

5. Conclusions

We have shown that both work in class and work with the Fluency system can improve pronunciation of “th” for learners of English. The consistent improvement we have observed in *all phonological contexts* encourages us to pursue the training of other sounds on the same basis. We should always keep in mind that this is an immersion situation and that results may vary in other contexts.

The use of the Fluency system in this and other experiments, such as those reported in [1], reinforce our belief that Fluency can be used as a testbed for questions about language learning in general as well as second language learning. For example, the balance between freedom given to the user and system control of events is presently under study.

We continue to design experiments that will lead to the discovery of more powerful techniques in pronunciation training and to the adaptation of such systems to the user.

Acknowledgements

The authors would like to thank Brian MacWhinney and Peggy Heidish for their suggestions in the design of this experiment. We would also like to thank Allie King for her invaluable help in grading the tests.

6. References

[1] Eskenazi, M., Ke, Y., Albornoz, J., Probst, K. (2000). The Fluency Pronunciation Trainer: Update and user issues, *Proc. INSTiL2000*, Dundee, Scotland.

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

[3] Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., Weintraub, M. (1990). Automatic evaluation and training in English pronunciation, *Proc. ICSLP*, Kobe, 1990.

[4] Neumeyer, L., Franco, H., Weintraub, M., Price, P. (1996) automatic text-independent pronunciation scoring of foreign language student speech, *Proc. ICSLP ’96*, Philadelphia.

[5] Witt, S., Young, S. (1998). Performance measures for phone-level pronunciation teaching in CALL, *Proc. STiLL*, Marhollmen, 1998.

[6] MacWhinney, B. (1995). Evaluating foreign language tutoring systems, in Holland, V., Kaplan, J., Sams, R, ed. *Intelligent language tutors: theory shaping technology*, Lawrence Erlbaum, Mahwah, NJ, 1995.

[7] Chapelle, C. (1997). CALL in the year 2000: Still in search of research paradigms?, *Language Learning and Technology*, 1(1):19-43, July 1997.

[8] Dunkel, P. (1991). The effectiveness research on computer-assisted instruction and computer-assisted language learning, in P. Dunkle, ed. *Computer-assisted language learning and testing*, p. 5-36, Newbury House, New York, 1991.

[9] Clark, John L.D., Swinton, Spencer S. (1979) An Exploration of Speaking Proficiency Measures in the TOEFL Context, *TOEFL research report 4,* Educational Testing Service, 1979.