Effects of Speech Recognition-based Pronunciation Feedback on Second-Language Pronunciation Ability

Kristin Precoda, Christine A. Halverson, and Horacio Franco

Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA
{precoda,krys,hef}@speech.sri.com

Abstract

This study's goal was to determine whether receiving a particular type of feedback on nativeness of second-language accent positively influenced pronunciation over time. Forty-five native speakers of American English of beginning to intermediate Spanish ability were randomly assigned to three groups. The first group was asked to practise Spanish using speech recognition-based software that provided scores of nativeness of pronunciation. The second group practised with software that was identical but with no feedback indicating pronunciation scores. The third group did not practise with the software. The subjects' speech was recorded at the beginning of the study and again after three weeks, and scores based on log posterior probabilities were calculated. The speech recognizer outputs these scores, which have been shown to correlate well with human listeners' nonnativeness judgements [1]. The two practice groups showed a small but statistically significant improvement over the control group. The two practice groups were also compared with each other.

1. Introduction

Language-learning software which includes speech recognition technology can give learners feedback on aspects of their oral performance. One aspect of particular interest to us is pronunciation, and speech technologists have worked on detecting specific pronunciation difficulties (e.g. [2,3,4,5]) and overall nonnativeness (e.g. [1,6,7]). In this study, we examine the other half of the feedback loop: that is, the influence the computer's feedback has on the learner's performance, in the context of one specific software implementation which offers feedback on the degree of nonnativeness of the user's pronunciation.

Figure 1 shows the user interface of the FreshTalk software, developed at SRI International and discussed in section 2.1. An alternate version of the software was also prepared which presented no pronunciation feedback to the user. Study participants were divided into three groups and were exposed either to FreshTalk with feedback, to FreshTalk without feedback, or to no software at all. There were four hypotheses to be tested:

1. Is practising with FreshTalk and receiving FreshTalk-style pronunciation feedback associated with an increase in log posterior probability scores, relative to practising with FreshTalk and not receiving pronunciation feedback?
2. Is practising with FreshTalk and receiving FreshTalk-style pronunciation feedback associated with an increase in speech rate, relative to practising with FreshTalk and not receiving pronunciation feedback?

The last two hypotheses, which were intended to be tested only if the first two showed no significant results, were:

3. Is practising with FreshTalk, with or without pronunciation feedback, associated with an increase in log posterior probability scores?
4. Is practising with FreshTalk, with or without pronunciation feedback, associated with an increase in speech rate?

These latter hypotheses would be tested by pooling the data from the two practice groups and comparing it against the data from the no-practice group.

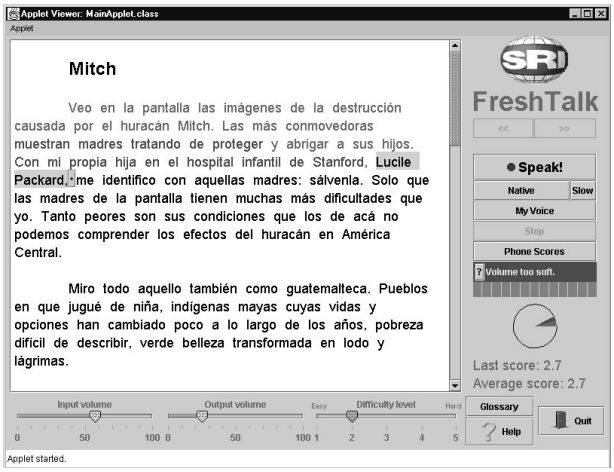

Figure 1. FreshTalk software user interface.

2. Experiment

2.1 Software

The software, FreshTalk, was designed at SRI International to demonstrate automatic pronunciation scoring on a reading task. Because of this focus, the user interface was not explicitly designed for long-term use. To read, the user selects a phrase, sentence, or paragraph anywhere in the presented text. She or he can play a native speaker's recording of the selection, either at normal speed or in a slow and careful style, and can record her or his voice reading the same selection for

playback. The user is free to repeat the selection as often as desired or to move on.

FreshTalk includes a speech recognizer which aligns the user's speech with a broad phonetic transcription of the selected text and calculates log posterior probability scores for the speech. The log posterior probability scores are mapped onto a scale from 1 to 6 such that the best correlation with human listeners' ratings of degree of nonnativeness is obtained. Given sufficient speech, the correlation between these scores and the average of those assigned by a panel of human listeners is comparable to that between two human listeners [1].

FreshTalk makes its pronunciation scores available to users in several ways. Among these are:

- The score of the last text selection is displayed on the screen, along with the average score over the entire session.
- The user can bring up a bar graph showing the total number of words receiving each score.
- After a phrase is read, it is displayed in green or red depending on whether its score is above or below a user-controlled threshold.
- A pie chart indicates visually what fraction of phrases received scores above or below the user-controlled threshold.

For this study, a modified version of the software was also produced which had the same functionality except that it did not provide any pronunciation scores. In this version, text selections were displayed in green after the user had read them; all other score-dependent features were removed from the program.

Other features of FreshTalk included a glossary of more difficult words or phrases. The software contained different texts for each practice session, on a variety of topics including current affairs, literature, biographical sketches, humour, and so on. These texts were written expressly for Spanish learners at a first-year university level. The intent was to minimise the difficulty of other aspects of the language so that users could focus their efforts on pronunciation.

2.2 Study length

The study was planned to last six weeks for each subject but was designed so that it could be terminated after three weeks if necessary. Due to a high dropout rate, only the data from the first three weeks will be discussed and analysed here.

2.3 Subjects

Forty-five subjects were recruited from first- or second-year Spanish classes at a local community college and through a variety of advertisements. All subjects were native speakers of American English with a beginning or intermediate knowledge of Spanish and reported normal or near-normal hearing. Subjects ranged in age from 19 to 54. Only five subjects were not actively studying Spanish, and all of those had studied it within the last five years.

Subjects were assigned to one of three groups, two of which practised their Spanish using FreshTalk and one of which did not. A few subjects who indicated they could not commit to the practice schedule were assigned to the no-practice group; other subjects were assigned randomly in such a way that the groups would be of approximately equal size. The two practice groups differed only in whether their version of the software included pronunciation scoring. They were not told that there were two versions.

The two practice groups were asked to use the software in three half-hour sessions per week for three weeks and were made aware that the software would log their practice time. One subject in the feedback group failed to practise at all and was reassigned to the no-practice group. The average length of time actually practised was 189 minutes for the feedback group and 191 minutes for the no-feedback group.

2.4 Recording sessions

Each subject was recorded upon beginning the study and again approximately three weeks later. Subjects were told that the purpose of the recording was to measure their pronunciation ability in Spanish. During the recording sessions, the subjects wore a small noise-cancelling headset microphone and read 53 sentences (3 for practice and 50 for later analysis) which were sampled at 16kHz and saved directly in digital format. The sentences, which were the same for all recording sessions, were chosen to include several sounds or sequences expected to be problematic and contained only relatively simple vocabulary.

A research assistant monitored microphone placement and audio quality throughout the session. The research assistant also asked the subject to repeat the sentence in cases where the subject stumbled, repeated a word, or read the wrong word entirely. The research assistants were instructed not to re-record a sentence on the basis of poor pronunciation: rather, their task was only to ensure that the audio quality was good and that the spoken words corresponded to the written words.

Most subjects completed the recording sessions in 20 to 30 minutes.

3. Analysis and results

3.1 Measures

User pronunciation performance was assessed in two ways. The first is an average of the log posterior probabilities for each phone (detailed in [1]) and measures the spectral match between the user's speech and statistical models trained on native speakers. This measure will be called the *posterior score*. The second is the speech rate, expressed in number of phones per unit time, using the speech recognizer's phonetic segmentation. Phones adjacent to silence were omitted because of potentially unreliable segmentations.

3.2 Covariates

Several covariates which we expected might be associated with changes in pronunciation over time were logged by the software. These included (a) number of utterances produced during practice sessions, (b) number of times a native-speaker example was played during practice sessions, and (c) number of times the user listened to her or his own voice during practice sessions. We discovered that no user ever listened to her or his own voice during practice sessions, so the last variable was dropped.

An analysis of covariance was planned to remove the effects of these uncontrolled variables before examining the effect of practising with or without pronunciation feedback. As the analysis of covariance requires that the treatments and covariates be independent, t-tests were used to check whether the two practice groups differed in total number of utterances produced or of native examples listened to [8]. No significant differences were found (unequal variance, two-sided t-test; df = 12.8 and 8.1, respectively; p = .91 and .75, respectively). Nor did the two practice groups differ in number of utterances or of native examples per minute of practice time (df = 9.0 and 9.5, respectively; p = .92 and .39, respectively). We concluded that the assumption of independence of treatments and covariates was reasonable.

In addition, linear regressions revealed no significant correlations between, on the one hand, number of utterances produced and number of native examples heard, and on the other hand, change in posterior score or change in speech rate, for either practice group. Possible relationships were also checked for visually. Plots showed no particular relationship between the covariates and change in posterior scores, or between number of utterances and change in speech rate. There did appear to be a small decreasing relation between use of native examples and change in speech rate, but no linearising transformation seemed necessary.

3.3 Results: comparison between practice groups

Linear regression of change in posterior score on practice group membership, number of utterances produced, number of native examples heard, and amount of practice time did not show any significant effect of practice group membership, singly or with other predictors.

Neither was there any significant effect of practice group membership on speech rate, whether or not number of utterances produced, number of native examples heard, and amount of practice time were taken into account. However, both number of native examples heard (p = .023) and practice time (p = .029) were significantly related to change in speech rate. The fitted model for change in speech rate was as follows:

$$R = -3.897 + 0.07P - 0.039N$$

where R is the change in speech rate, P is the number of minutes of practice time, and N is the number of times a native example was listened to (multiple $r^2 = .339$). A negative R indicates that the subject spoke more quickly the second time she or he was recorded, after practising for several weeks with FreshTalk. It is interesting to note the signs of the coefficients: a longer time spent practising is associated with slower speech, while listening to native examples is associated with an increase in speech rate. It is possible that increased practice time leads to slower speech through increased attention to pronunciation.

3.4 Results: comparison between practice and no-practice groups

Visual inspection showed no clear violations of normality for either change in posterior scores or change in speech rate, for either the practice groups or the no-practice group. Therefore, t-tests were used to check for differences between practisers and nonpractisers. A significant difference in change in posterior scores was found (p = .0034). However, the estimated mean change for practisers was a small increase, while for nonpractisers, there was a mean decrease of similar size. This apparent decrease is difficult to explain. It may indicate a change in the subjects' attitude toward the recording procedure, or simply an insufficient sample size.

No significant difference was found between practisers and nonpractisers for change in speech rate (p = .72).

3.5 Results: retention rate

As expected, some subjects participated in the first recording session but did not return for the second session three weeks later. The retention rates for the three groups are shown in the following table. Subjects from the community college are tabulated separately from other subjects because they had easy access to an on-campus computer for practice sessions and might be expected to be relatively homogeneous in retention rate. Other subjects had to come to our lab to practise.

Table 1. Retention rates: Number of subjects who returned for the second recording session, out of those completing the first recording session

Group	Community college subjects	Other subjects	Total
Practice, with feedback	9/12	3/3	12/15
Practice, without feedback	6/11	0/3	6/14
No practice	9/10	5/6	14/16

It appears that practising with pronunciation feedback led to a substantially higher retention rate than practising without pronunciation feedback, though no post-hoc test of significance is possible. Interestingly, of the 8 out of 14 no-feedback subjects who did not return for a second recording session, only two ever practised with

FreshTalk, according to the software logs. The others did not begin practising and then lose interest; rather, they were introduced to the no-feedback version of the software during their first session at SRI and apparently never used it on their own.

4. Summary and discussion

The above results can be summarised as follows:

- The two practice groups were not found to differ in either how many utterances they produced or how many times they listened to native speaker examples, either in total or per minute of practice.
- No significant difference between the two practice groups was found in either change in posterior scores or change in speech rate.
- For the practice groups, increasing speech rate was significantly correlated with increasing use of native examples, and decreasing practice time.
- Subjects who practised with the software achieved a small but significant increase in posterior scores, relative to subjects who did not. No difference between practising and nonpractising subjects was found in change in speech rate.
- The version of the software with pronunciation feedback appears to be associated with a greater willingness to use the software.

There are many possible reasons why there is no apparent difference between the performances of the practice group receiving feedback and the practice group not receiving feedback. One, of course, is that the feedback is not in itself useful. However, it is also possible that the same feedback might have an effect over a longer time period; or that the differences between individuals overwhelm the differences between groups as small as these; or that posterior scores are an inappropriate measure of actual change; or that the particular ways in which feedback is given are not well designed. There was some evidence in subjects' comments that the presentation of feedback could be improved upon. There may also be interactions between feedback and characteristics of the language under study and the native language, or between feedback and the initial level of skill of the language learners. Informal observation during the present study, for example, suggested that an explicit focus on Spanish letter-to-sound rules might have been helpful in improving the pronunciations of some subjects.

Several subjects in the practice groups reported greater confidence and ease of oral production in class after participating in this study.

It is interesting to compare the present results with those in [9]. In that study, students used a computer-based system primarily to practise controlling their pitch and loudness, and responded enthusiastically to the software yet did not show any greater gains than students spending an equivalent amount of time in traditional individualised instruction. In our study, while subjects appeared to appreciate receiving automatic feedback, they did not perform better than subjects using the version of the software without feedback.

Perhaps the most important conclusion to be drawn is that the design of the user interface deserves as serious attention as the underlying technology.

Acknowledgements

Our deepest thanks are due the participants in this study. We also thank Dr. Carlos Lopez of Menlo College and Dr. Maria Cristina Urruela and Al Moreno of Stanford University, and our research assistants, Jane Ancheta, Inna Matov, and Colleen Richey. This work was supported by the U.S. Government under DARPA Agreement DASW01-96-3-0001. The views expressed here do not necessarily reflect those of the Government.

References

[1] Neumeyer L, Franco H, Digalakis V and Weintraub M (2000). Automatic Scoring of Pronunciation Quality, *Speech Communication*, 30: 83-93.

[2] Eskenazi M (1996). Detection of Foreign Speakers' Pronunciation Errors for Second Language Training: Preliminary Results, *Proc. International Conference on Spoken Language Processing 96*, Philadelphia, PA, 1465-1468.

[3] Franco H, Neumeyer L, Ramos M and Bratt H (1999). Automatic Detection of Phone-Level Mispronunciation for Language Learning, *Proc. Eurospeech 99*, Budapest, Hungary, 851-854.

[4] Herron D, Menzel W, Atwell E, Bisiani R, Daneluzzi F, Morton R and Schmidt J (1999). Automatic Localization and Diagnosis of Pronunciation Errors for Second-Language Learners of English, *Proc. Eurospeech 99*, Budapest, Hungary, 855-858.

[5] Witt S and Young S (1997). Language Learning Based on Non-Native Speech Recognition, *Proc. Eurospeech 97*, Rhodes, Greece, 633-636.

[6] Cucchiarini C, Strik H and Boves L (1998). Automatic Pronunciation Grading for Dutch, *Proc. ESCA Workshop on Speech Technology in Language Learning (STiLL 98)*, Marholmen, Sweden, 95-98.

[7] Franco H, Neumeyer L, Digalakis V and Ronen O (2000). Combination of Machine Scores for Automatic Grading of Pronunciation Quality, *Speech Communication*, 30: 121-130.

[8] Cochran WG and Cox GM (1957). *Experimental Designs*, 2nd ed., Wiley, New York.

[9] Stenson N, Downing B, Smith J and Smith K (1992). The Effectiveness of Computer-Assisted Pronunciation Training, *CALICO Journal*, 9/4: 5-18.