

SPEECH RECOGNITION FOR LANGUAGE TEACHING AND EVALUATING: A STUDY OF EXISTING COMMERCIAL PRODUCTS

Rebecca Hincks

Centre for Speech Technology
Department of Speech, Music and Hearing
KTH, Stockholm, Sweden
hincks@speech.kth.se

ABSTRACT

Educators and researchers in the acquisition of L2 phonology have called for empirical assessment of the progress students make after using new methods for learning [1]. This study investigated whether unlimited access to a speech-recognition-based language learning program would improve the general goodness of pronunciation of a group of middle-aged immigrant professionals studying English in Sweden. Eleven students were given a copy of the program *Talk to Me* by Auralog as a supplement to a 200-hour course in Technical English, and were encouraged to practice on their home computers. Their development in spoken English was compared with a control group of fifteen students who did not receive software. *Talk to Me* uses speech recognition to provide conversational practice, visual feedback on prosody and scoring of pronunciation. A significant limitation of commercial systems currently available is their inability to diagnose specific articulatory problems. However, in this course students also met at regular intervals with a pronunciation tutor who could steer them in the right direction for finding the most important sections to practice for their particular problems. Students reported high satisfaction with the software and used it for an average of 12.5 hours. Students were pre- and post-tested with the automatic PhonePass SET-10 test from Ordinate Corp. Results indicate that practice with the program was beneficial to those students who began the course with a strong foreign accent but that students who began the course with intermediate pronunciation did not show the same improvement.

1. INTRODUCTION

1.1. Speech recognition for teaching and evaluating

With proper adaptation, speech technology allows beginning language students to practice spoken language outside the classroom. Dialogue-based software using fixed-response ASR lets learners have a simulated conversation with a computer. Practicing with such programs should help students improve fluency and confidence. Furthermore, the software can provide individual feedback on pronunciation, which is something that is often lacking in the language classroom. Algorithms calculate by how much a given pronunciation has deviated from a model, and give a score on phonetic accuracy. Visual representation of

waveforms and pitch curves aid students in perceiving how their prosody deviates from a model. Computer animations illustrate the inside of the vocal tract and the movements necessary for proper articulation of sounds. Features such as these should be an important aid to teachers who do not have the resources to instruct students individually in pronunciation.

Morley [1] expresses a “need for controlled studies of changes in learner pronunciation patterns as the result of specific instructional procedures”. The research reported on here was a study of whether an existing commercial program using speech recognition could help improve student pronunciation. Students in one section of a course in Technical English were offered their own copy of a leading program for practicing spoken English. The development in their pronunciation was compared with that of students in another section of the course. The instructional procedure to be evaluated was thus the use of a particular pronunciation program as supplemental courseware.

A similar study was carried out by [2], who asked beginning and intermediate learners of Spanish to practice with a program of the researchers’ own design, *FreshTalk*. There were three groups of students, two using different forms of the software, and one that had no software at all. Students were recorded at the beginning and end of the three-week trial, and practiced an average of three hours. A conclusion these researchers found was that the interface design was important in keeping the students using the software. The study suffered from a high drop-out rate. Language learning software is of course of no benefit at all if it is unused, and getting students to devote their own time to extra practice is easier if the software is pleasant to use. This is justification for using commercial products where resources have been put into the sensory stimulation provided by musical and photographic material, as well as a cohesive and amusing interface.

An obstacle in testing pronunciation is determining a practical method for evaluation. Human judgment is time-consuming and it is difficult for raters to be consistent. In this study, a fully automatic measure of pronunciation was used for pre- and post-testing the control and experimental groups. This was the PhonePass SET-10 test from Ordinate Corp [3], a ten-minute test of spoken English administered over the telephone [4]. The test uses speech recognition to assess the correctness of student responses and also gives scores in pronunciation and fluency. The pronunciation score for each student is the basis for the analysis reported in this paper. An evaluation of the successfulness of using PhonePass for scoring pronunciation is described in [5].

1.2. Software used for this project

One company that has successfully marketed language learning programs that use ASR is the French company Auralog [6]. Auralog thoroughly integrated the technology into its *Talk to Me* language series, which is the product we chose for our study. The core of the software consists of six dialogues, where the user utters one of three responses. Each dialogue consists of thirty question-and-answer screens with accompanying photographic illustrations and occasionally music and video clips as well. The act of choosing a response initiates a degree of spontaneity into the 'conversation' and hopefully allows more natural language than would be enabled by just reading one specific response.

While the dialogues in practice communication skills, more specific pronunciation training is carried out at sentence, word, or phoneme level. At phoneme level, users are shown 3D animations of a sagittal section, showing how the sound is articulated. At word and sentence levels, each response from the dialogues is practiced individually. Users compare the waveform and pitch curve of their own production with that of the model speaker. A score for the production is given, on a scale from one to seven. If the program has found particular difficulties recognizing a specific word in the phrase, that word is highlighted in the text screen, to help the student notice what sounds he needs to work on. The user's responses are recorded and she can listen to any of them at any time. The program is thus a development of a record/playback model, with the added input of feedback in the form of a score from the system, extraction of the most serious deviation from the models, and the visual comparison of speech signals.

The user can adjust the levels of different settings in the software. The speech can be slowed down to allow easier comprehension. Most important, however, is that the difficulty level of the speech recognition can be adjusted to require a looser or tighter match to the underlying models. For example, at the lowest level of difficulty, the system recognized and accepted the word *villa* pronounced as 'willa'. This pronunciation was rejected and the phrase unrecognized at higher levels of difficulty. This allows users to challenge themselves by setting harder levels.

An important limitation of this software, like other commercial systems currently available, is its inability to diagnose specific articulatory problems and give corrective rather than evaluative feedback. In our study, however, students also met with a tutor who helped them diagnose and correct their problems.

2. METHOD

2.1. Framework

This study involved two groups of students, the control group taking a course in the fall of 2000, and the experimental group taking the same course the following term, spring 2001. The experimental group received *Talk to Me English (1)* as supplemental courseware. The course was a 200-hour, ten-week course in Technical English for Immigrants offered at the Royal Institute of Technology (KTH) in Stockholm. One requirement of the course was five hours of individualized, tutor- and computer-assisted pronunciation tutoring. In the fall term of 2000, students followed the normal course plan and were tested for the purposes of future comparison. Their five hours of pronunciation tutoring were assisted by a software program that helped the students learn IPA notation (*Skytalk*) so that they could use dictionaries more effectively. They did not receive their own pronunciation program. In the spring term of 2001, students were offered the opportunity

to trade one hour's tutoring for a copy of *Talk to Me* for use on their home computers. They still received four hours of tutoring, using *Talk to Me* instead of the IPA program. The course content was generally the same for the two groups, but the teachers were different.

2.2. Subjects

The students in the course were middle-aged engineers from different language backgrounds. The course was funded by the Stockholm County Employment Bureau and was intended to encompass full-time, paid activity for the students. The students were admitted to the course on the basis of a placement test, but possessed varying skills in English ranging from advanced beginner to upper intermediate. In the control group there were fifteen students (two female), who were native speakers of Farsi (3), Spanish (2), Arabic (2), Azerbaijani (2), Kurdish, Polish, Russian, Armenian, Tigrinya and Romanian. Their average age was 42.

The pool of students in the spring course of 2001 (the experimental group) was thirteen. Eleven of them accepted the offer to trade an hour of tutoring for a program, and were given a CD-ROM copy of *Talk to Me*. Ten of these were able to successfully install the program, and nine of them practiced with it. This group consisted of seven males and two females, age 47 in average, who were native speakers of Farsi (2), Polish (2), Hungarian, Romanian, Russian, Somali and Arabic.

2.3. Use

Students were asked to keep a log of how many hours they used the program, but were not assigned practice as homework in any strict sense. The number of hours of use at home ranged very widely, from 2 to 48, with a mean of 12.5, sd 15. The students also used the program for four hours with a pronunciation tutor.

2.4. Pre- and post-testing

Students' production of spoken English was evaluated both at the beginning and end of the course by completion of the ten-minute PhonePass test. Test papers were purchased from Ordinate Corp. Though the PhonePass test can be administered from any location, our students made their phone calls from a soundproof room at KTH and were digitally recorded using 16-bit sampling. The incoming signal from the telephone line was also digitally recorded.

3. RESULTS

3.1. Satisfaction

At the end of the course, the nine students using *Talk to Me* filled out a questionnaire about their attitudes toward the program. They reported that the program was fun to use and thought it benefited their English. Most reported that they were not able to use the program as much as they had hoped, due to lack of time partially caused by the amount of assignments in other components of the language course.

3.2. Pronunciation

Taken as wholes, neither the control nor the experimental group showed significant improvement in pronunciation.

The PhonePass test provides six scores: an overall score and five sub-scores. This paper looks at the sub-score for

pronunciation, which reflects how the examinee differed from the acoustic models in the speech recognition system. It is thus a reflection of phonetic accuracy. The minimum score a test-taker can receive is 2 and the highest is 8. Ordinate reports that a difference in scores of more than 0.2 can be considered significant (Bernstein, personal communication). Neither the control group nor the experimental group produced changes in mean group scores from the pre-test to the post-test that exceeded this amount. Figure 1 shows the pre- and post-testing results for the control group, and Figure 2 shows the same for the experimental group.

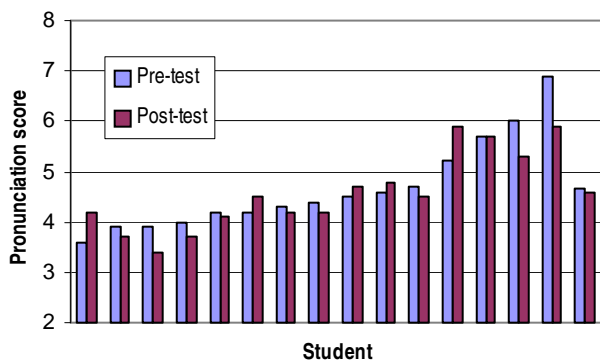


Figure 1. Control group: Pre- and post-test scores in Pronunciation from PhonePass test.

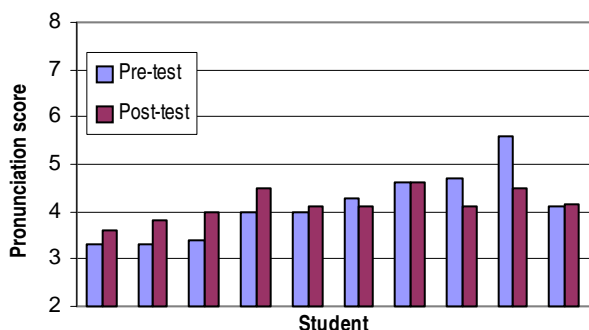


Figure 2. Experimental group: Pre- and post-test scores in Pronunciation from PhonePass test.

The last column in both figures shows the mean scores of the tests. The control group shows an insignificant decline from 4.7 to 4.6, and the experimental group no change at all at 4.1.

Taken as individuals, an important division emerges when students are grouped according to their proficiency level. Students with poor pronunciation to begin with, i.e. those with a score of less than 4.2, a group defined by Ordinate as having an ‘intrusive’ foreign accent, appear in the leftmost columns of Figures 1 & 2. This group improved much more in the experimental group than in the control group. The mean score of the five weakest experimental students showed a significant improvement of 0.4, while the six students in the control group with scores less than 4.2 decreased their mean score very slightly. Figure 3 presents the

change in pronunciation scores for students, broken down into the degree of accentedness with which they began the course. The only group showing improvement is the strongly accented students in the experimental group.

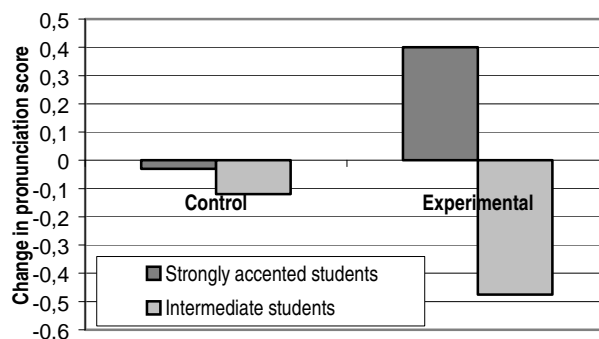


Figure 3. Change in pronunciation score for students according to beginning level. Weak students improved significantly in the experimental group, while all other groups showed no improvement.

There was no clear relationship between the amount of time spent using the software and degree of improvement. The four students who used the program the least, however, showed the most improvement, as shown in Figure 4.

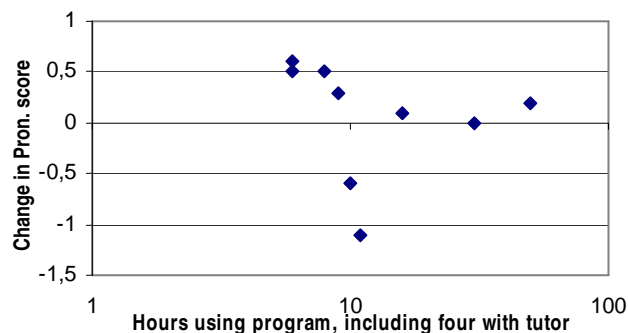


Figure 4. Relationship between time spent using program and change in pronunciation score on PhonePass test.

4. DISCUSSION

4.1. Teaching

The small size and diverse skills of the experimental group make it difficult to draw conclusions that are statistically significant. Furthermore, it may have been unrealistic to expect any improvement in pronunciation for these students in this time frame. Acton [7] reports some improvement for only 50% of a group of foreign professionals who took a course in English pronunciation that involved 120 hours of work on speech only. Our course in Technical English taught all aspects of the language, with pronunciation specifically focused on for an average of only 12.5 hours.

Talk to Me provides language learning potential by highlighting the ‘worst’ word in an utterance and giving scores on

pronunciation. Still, users need guidance in finding the necessary information to correct their mistakes. An example of how these activities work in practice is the student who came to his tutor after using the software and mentioned that he had noticed that the computer responded negatively to all his pronunciations of words beginning with /p/. The tutor explained that he was not aspirating the sound. An animated video of the articulation of /p/ was available to the student, and ideally he would have been able to use it to diagnose and solve his problem. Still, the software was at least able to make the student notice that he had problems with a particular phoneme. Having the student actively engaged in the diagnosis of his pronunciation difficulties is pedagogically desirable. The ideal automatic system would also point the student in the direction of the appropriate remedial activities. The fact that the four students showing the greatest improvement had used the program the least on their own could indicate that it was important that a large proportion of the learning time was spent in conjunction with the human tutor.

The results of this study show just how important it is that the pedagogical tools be appropriate to a given student's level of development in the target language. The learner fit of this program seemed to be better for students with a strong foreign accent than for those with an intermediate accent. Auralog makes an attempt to make the software adaptable to a range of users by making the recognition requirements adjustable, demanding a more exact match of the model phrase at higher levels. This should allow students who are already at an intermediate level of proficiency to challenge themselves by imitating the models, and in its instruction booklet, Auralog claims that practice in imitation will lead to better pronunciation. The failure of our better students to improve in their pronunciation is evidence that mimicry does not necessarily improve pronunciation. In fact, these results support the advice given by [1] that imitative speaking practice be used only for beginning students of a language.

4.2. Evaluating

A potential problem with the methodology used in this study is the fit between the varieties of English used for assessing and for teaching. The PhonePass test is designed to assess how well candidates will do in a North American environment. The speech corpus used to train the hidden Markov models used by the speech recognition is of North American English. Ordinate has accommodated candidates who have learned British English by checking responses at the lexical level with British linguists. The programs from Auralog, on the other hand, used a fair amount of British English models. Our students were thus being taught both British and American English (of the five teachers in the course, three spoke Am.E and two Br.E) but were being evaluated on only American English. This may not have been a problem for the more strongly accented students, but perhaps it was for the better students who were ready to move in the direction of native-like pronunciation.

It was naturally discouraging for two students who had devoted an extraordinary amount of time to extra pronunciation work to see that they received no positive recognition for their efforts from the PhonePass test. One of these students received a substantially lower score for reading fluency (a temporal measure) indicating that she had slowed down her speech as she attempted to produce more accurate segments. Similar results were seen in the study carried out by Precoda et al. [2] who found a negative correlation between practice with language learning software and change in speaking rate, indicating that increased attention to

pronunciation could slow down speech. In a study comparing human judgements of spoken Dutch with automatic evaluation, Cucchiariini et al. [8] found that the best automatic predictors of overall pronunciation ability were temporal measures such as rate of speech and total duration. However, these same measures could be misleading if ASR is applied to measuring *development* in pronunciation.

5. CONCLUSION

Extra pronunciation training using ASR-based language learning software did not demonstrably improve the mean pronunciation abilities of a heterogeneous group of middle-aged students. However, results from the PhonePass test indicate that use of the program was beneficial for the students who began the course with an 'intrusive' foreign accent. A comparable group did not improve if they were in the control group, despite the five hours of pronunciation tutoring they received.

6. ACKNOWLEDGMENTS

The author would like to thank David House, Björn Granström, Martin Dahlqvist, Margaretha Andolf, and the students and teachers at the Unit for Language and Communication at KTH for their help and participation in this study.

7. REFERENCES

- [1] Morley, J. "The Pronunciation Component in Teaching English to Speakers of Other Languages." *TESOL Quarterly* 25 (3) 481-520, 1991.
- [2] Precoda, K; Halverson, C. & Franco, H. "Effects of Speech Recognition-based Pronunciation Feedback of Second-Language Pronunciation Ability" In *Proceedings of InStil 2000* Dundee, Scotland, 2000., pp.102-105.
- [3] www.ordinate.com
- [4] Townshend, B; Bernstein, J; Todic, O; & Warren, E. "Estimation of spoken language proficiency". In *Proceedings of ESCA Workshop on Speech Technology in Language Learning (StiLL 98)*. Stockholm: KTH Department of Speech, Music and Hearing, 1998, pp.179-182.
- [5] Hincks, R. "Using speech recognition to evaluate skills in spoken English". in *Papers from Fonetik 2001*, Lund University Department of Linguistics, 2001, pp. 58-61.
- [6] www.auralog.com
- [7] Acton, W. "Changing Fossilized Pronunciation." *TESOL Quarterly* 18 (1), 71-85, 1984.
- [8] Cucchiariini, C; Strik, H; & Boves, L. "Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms" in *Speech Communication* 30, 109-119, 2000.