ReCALL **15** (1):3-20. © 2003 Cambridge University Press
DOI: 10.1017/S0958344003000211 Printed in the United Kingdom

Speech technologies for pronunciation feedback and evaluation

REBECCA HINCKS
Centre for Speech Technology, Dept of Speech, Music and Hearing, Kungliga Tekniska Högskolan (KTH), Drottning Kristinasväg 31, Stockholm, Sweden
(*Email: 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 (Chun, 1998, Morley, 1991). The present study investigated whether unlimited access to a speech-recognition-based language-learning program would improve the general standard 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* from Auralog as a supplement to a 200-hour course in Technical English, and were encouraged to practise on their home computers. Their development in spoken English was compared with a control group of fifteen students who did not use the program. The program is evaluated in this paper according to Chapelle's (2001) six criteria for CALL assessment. Since objective human ratings of pronunciation are costly and can be unreliable, our 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 was of limited value for students who began the course with better pronunciation. The paper begins with an overview of the state of the art of using speech recognition in L2 applications.

1 Introduction

Students of a foreign language must master the four main language skills: reading, listening, writing and speaking. The first three of these areas lend themselves naturally to CALL applications. Computers excel in particular as a tool for providing means for delivery and practice of writing. From process writing to chat, computers naturally channel communication for keyboard generation. Computers are, however, less versatile as a training medium for our primary means of communication: speech. The challenges for adapting digital systems to handle speech range from the technical to the linguistic. Technically, handling speech requires larger data storage and transfer capacity, additional equipment such as adequate microphones, and access to computers in quiet environments. Linguistically, speech processing is challenged by the natural variability of the speech signal and the complex chain of events required to generate a suitable response to a spoken utterance.

https://doi.org/10.1017/S0958344003000211 Published online by Cambridge University Press

Despite these challenges, engineers in the field of speech research have long hoped to be able to use their technology to help language learners, and language pedagogues have searched for ways in which the spoken language could be better supported by CALL materials. This paper surveys possible applications and existing constraints for using speech processing for feedback and evaluation of pronunciation. The paper also presents the results of a study designed to test the value of using state-of-the-art spoken language processing to provide automatic feedback and evaluation of the pronunciation of English as a second language.

Pronunciation is an area of language teaching and learning that is sometimes overlooked. In the nineteen-fifties and sixties, audio-lingual teachers drilled pronunciation with in-class repetition and recitation. In the seventies and eighties, communicative language teachers dismissed pronunciation as an 'extraneous' feature of the language (Brown & Yule, 1983:53). In many language classrooms, pronunciation ceased to be taught (Derwing & Rossiter, 2002:160). Teachers were unsure as to the best method for teaching it, and research showed that, just as in many aspects of language learning, there were large individual variations among learners as to what method was most effective (MacDonald, Yule & Powers, 1994).

Pronunciation training is well suited to be carried on outside the classroom for a number of reasons. In a monolingual classroom, though learners may share the same difficulties with the target language, the pronunciation errors they make may not impede their intelligibility to each other (Jenkins, 2000:58). When there is no penalty (in terms of lack of understanding) for inaccurate pronunciation, there is a risk of fossilization of mispronunciations. In a mixed-language classroom, on the other hand, learners do not necessarily share the same specific training needs. This makes it difficult for the teacher to devise in-class exercises that are meaningful for a majority of the students. In any classroom, some learners are uncomfortable when the spotlight is focused on their own production, and some teachers are unsure of how to give the necessary corrective feedback. In addition, many English teachers around the world today have themselves not had the opportunity to develop pronunciation suitable for use as a teaching model. Here is an area in need of effective CALL products.

The language laboratory has traditionally provided a means for necessary outside-the-classroom training, and as language labs around the world go digital, the opportunity arises for the development of software that can provide more than the traditional record-and-play-back model of spoken language training. Ideally such software should also be available for use on home computers, avoiding the inconvenience of travelling to a lab. Language lab practice is most beneficially done in association with the language teacher, who can listen to student production individually, giving feedback and comments. This is of course time-consuming, and when educational institutions around the world are forced to cut costs, teacher involvement in pronunciation training can be sacrificed. Using some form of automatic feedback would be a solution to this problem.

2 Spoken language processing for CALL: the state of the art

This section briefly surveys some of the ways speech technologies are currently used in L2 applications. Signal analysis software has long been used for teaching intonation; speech recognition is now used for practising and evaluating second languages.

https://doi.org/10.1017/S0958344003000211 Published online by Cambridge University Press

2.1 Computer-assisted feedback using signal analysis software

A teacher of pronunciation needs to be able to help a student with both perception and production of the target sounds in the L2. For decades now, teachers have supported their instruction by using signal analysis software to visualize the speech signal, with demonstrable benefits. This work is reviewed by Chun (1998). Studies have shown that presenting displays of pitch contours improves both perception and production of intonation. Groundbreaking work was done in the late 70s in the Netherlands by de Bot (e.g. de Bot & Mailfert, 1982), and in Sweden by Bannert (1979), both of whom showed that even limited training with audio-visual feedback of prosody was more beneficial than audio feedback alone. A similar line of investigation was later carried out by Molholt (1988) on Chinese-speaking learners of English, and by Öster (1999) on immigrants to Sweden. Recently, Hardison (2002) has expanded this work to show that audio-visually trained learners of French not only improved their prosody but also their segmental accuracy. These studies have been conducted in situations where there was a teacher available for guidance and interpretation. Because most language learners have little knowledge of phonetics, expert assistance is required for learners to extract value from pitch displays. Essential feedback is provided by the human, with the computer as a mere tool for visualization.

2.2 Constraints on using automatic speech recognition in language learning

The signal analysis software used for displaying pitch contours is a mature technology that has been functioning reasonably well for many years. A newer technology used in CALL for the last decade is automatic speech recognition, ASR. ASR has held the tantalizing promise of enabling a truly communicative, feedback-providing framework for CALL, by letting learners 'converse' with a computer. However, significant advances in natural language processing and computational power are necessary before native speakers can converse with a computer about anything beyond the constraints of limited domains. These challenges are multiplied for the prospect of accented users using speech recognition, since their pronunciations cannot be represented in a general language database without diluting the precision of the recognition (Egan & LaRocca, 2000).

The basis of ASR technology is the probabilistic comparison between the signals received by the system and what is known about the phonemes of a language as represented in a database containing recordings of hundreds of native speakers of the language. Because of ASR's mathematical nature, numerical scores can be derived representing the deviation between a signal and an acoustic model of the phoneme it is hypothesized to represent. These scores can then be given to the learner as a type of feedback measuring a quantifiable distance from a target phoneme. However, it is not possible with current technology to say in what way the signal has deviated from the model, and this means that feedback is not corrective or constructive, but merely a sort of evaluation of the signal. Some current research looks at the creation of ASR databases specific to a learner's L1, and from these anticipates mistakes the user is likely to make (Delmonte 2000; Menzel, Herron, Bonaventura & Morton 2000; Tomokiyo, 2000; Minematsu, Kurata & Hirose 2002). By expecting a user to, for example, substitute an

https://doi.org/10.1017/S0958344003000211 Published online by Cambridge University Press

unvoiced sound for a voiced sound, the proper feedback message can be prepared for delivery when that voiced sound receives a low score.

Another constraint involved in using ASR in language learning is that speech recognition systems at present are very poor at handling information contained in the speaker's prosody. In order to recognize the words of an utterance, the recognition engine must ignore the variations of pitch, tempo and duration that naturally appear in utterances by different speakers and even within an individual speaker's various productions. This means that ASR can give feedback on the segmental level, but not on the prosodic. Unfortunately for CALL developers, prosodic features are often those that need the most practice from language learners (Anderson-Hsieh, Johnson & Koehler, 1992). ASR can, however, be successfully used to measure the speed at which a learner speaks, a type of fluency measure. Rate of speech has been shown to correlate with speaker proficiency (Cucchiarini, Strik & Boves 2000). Thus, the best prosodic application of ASR to date is in assessment of speaker fluency.

Speaker age and gender is another issue that must be taken into consideration when using speech recognition. Many language learners are children, but the higher frequencies of their voices makes their speech unsuitable for recognition in systems based on databases of adult speech. Special programs need to be created for them. Ideally, the system should be sensitive to the sex of the user, so that users model their utterances on those of speakers of the same sex. Work on allowing users to pick their own model speaking voices has been carried out by Probst, Ke & Eskenazi (2002).

Finally, it must be acknowledged that the collection of speech databases and the creation of speech recognition engines is an expensive and time-consuming process. This means that commercial organizations may find it not viable to invest in developing products for less-spoken languages.

2.3 Strategies for using automatic speech recognition in language learning

Despite the above-mentioned limitations to ASR applications in language learning, products using the technology arrived on the commercial market in the mid 1990s. Since the starting point for the research reported on in this paper was the necessity of using pre-existing ASR-based software for teaching oral skills in English, an investigation was conducted into the systems available (in the fall of 2000) at the commercial and research levels. The product then known as *Talk to Me* by the French company Auralog was chosen for testing.[1]

2.3.1 Talk to Me

Talk to Me was chosen because we judged it to attractively apply pedagogical goals within the constraints of existing ASR technology. Auralog originated as a speech technology company, and moved into language learning as a promising application. Instead of adding a speech-processing component to an already existing program, which was the strategy taken by competing language-learning companies, Auralog built up the program

1 Auralog has more recently integrated this software as the oral practice componant of its complete language-learning course *Tell Me More* (www.auralog.com).

https://doi.org/10.1017/S0958344003000211 Published online by Cambridge University Press

around what was possible from an engineering standpoint. While this approach could be criticized as being technology-driven, the result has been a series of products that have been commercially successful.

The core of the software consists of six dialogue sequences, where the program asks a question to which the user responds by uttering one of three answers. If the answer is recognized by the program, the dialogue moves along to the next stage of the 'conversation'. The act of choosing a response initiates a degree of spontaneity into the dialogue and hopefully allows more natural language than would be enabled by just reading one specific response. The performance of the speech recognition is at the same time facilitated by having to choose between only three possible answers (an approach also taken by other designers: for example, Holland, Kaplan & Sobol, 1999: 346).

Each dialogue consists of thirty question-and-answer screens with accompanying photographic illustrations and occasionally music and video clips as well. The visual material seems to have been chosen for its potential to give pleasant sensory stimulation and enrichment; the photographs are of beautiful beaches, adorable children, delicious food, etc. Interface design has been shown to be important in keeping the students using the software (Precoda, Halverson & Franco, 2000).

While the dialogues in *Talk to Me* practise general communication skills, more specific pronunciation training is carried out at sentence, word, or phoneme level. At phoneme level, users are shown animations of the vocal tract, showing how the phoneme is articulated. At word and sentence levels, each response from the dialogues is practised individually. Here the speech recognition is augmented by signal analysis to give visual feedback of the intonation of the utterances. Users compare the waveform and pitch contour 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. The user's responses are recorded and can be played back. 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 display of wave form and pitch curve.

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* with an initial /w/-sound. This pronunciation was rejected and the phrase un-recognized at higher levels of difficulty. This allows users to challenge themselves by setting more demanding levels.

2.3.2 Other products and projects using ASR in language training

In this section we will look briefly at some of the other work where ASR has been applied with pedagogical goals. The other early commercial operation involved in using speech recognition for L2 applications was Syracuse Language Systems, which produced two programs with ASR: *TriplePlayPlus* and *Accent Coach*. In 2002, these products have been absorbed into software packages from other language learning companies. New companies continue to enter the market as recognition engines are

https://doi.org/10.1017/S0958344003000211 Published online by Cambridge University Press

made available to software developers. For example, *EduSpeak*, (www.eduspeak.com) an SRI spin-off, sells recognition engines that have been adapted for non-native speech. They have collected speech databases of accented English and Spanish and thereby improved the accuracy of the recognition for non-native speakers (Franco, Abrash, Precoda, Bratt, Rao, Butzberger, Rossier, & Cesari, 2001).

The most widely known application for ASR is in dictation systems. Usually these systems are speaker-dependent; that is, trained to recognize the speech of one individual. A few researchers have been inspired to try these dictation systems on language learners, with little success. Coniam (1999) and Derwing, Munro & Carbonaro (2000) looked at the success with which foreign-accented speakers of English could use the commercially available dictation program *NaturallySpeaking* from Dragon Systems. Coniam had ten competent Cantonese-accented speakers of English train the dictation system on their voices and then analyzed the accuracy with which the system transcribed a read text. Predictably, the software was significantly worse at recognizing foreign-accented speech than native speech. Derwing *et al*. compared machine recognition with human intelligibility scores derived by transcribing recorded utterances. Like Coniam, they found that competent non-native speech was recognized much less accurately than native speech; moreover, they found a discrepancy between errors perceived by humans and the misrecognitions of the dictation software. The problems the dictation systems encountered did not reveal a human-like pattern. The implication is that CALL products using ASR could teach users more about speaking with a machine than about speaking with a person.

Dictation software has, however, not been designed with CALL applications in mind. ASR for non-native speech needs to be adapted so that the underlying phonetic models encompass a wider variety of possible productions. This was one of the goals of the EU-funded ISLE project, which ran from 1998-2000 (Menzel, Herron *et al*. 2001). Work on this project aimed at modeling Italian- and German-accented English to help intermediate learners. As a first step it was necessary to create a corpus of phonetically transcribed accented speech. This turned out to be very difficult to do reliably, given the limited annotation system required by the recognition engine (Bonaventura, Howarth & Menzel, 2000.) As a result, the final product performed poorly (Menzel *et al*. 2000). This is unfortunate as the designers had hoped to be able to provide corrective, rather than merely evaluative, feedback.

2.4 Evaluating language with automatic speech recognition

An obstacle in testing pronunciation is determining a practical method for evaluation. Human judgment is time-consuming and it can be difficult for raters to be consistent. However, as we have seen, speech recognition can be used for speech evaluation. While it is not possible for ASR to say *in what way* a sound has differed from an underlying model, it can be used to say *how much* a sound has deviated from an underlying model. In this section we will look at some of the ways ASR is being used to evaluate spoken language.

2.4.1 The PhonePass test

The PhonePass test from Ordinate Corporation (www.ordinate.com) is designed as a simple way for organizations to test the English skills of potential employees or stu-

https://doi.org/10.1017/S0958344003000211 Published online by Cambridge University Press

dents. In this sense it is a sort of automatic substitute for an oral proficiency test. Our students were pre- and post-tested using the PhonePass SET-10 test (e.g. Townshend, Bernstein, Todic & Warren, 1998), a ten-minute test of spoken English administered over the telephone. The test uses speech recognition to assess the correctness of student responses and also gives scores in pronunciation and fluency.

To administer the test, an organization purchases test papers from Ordinate Corp. To take the test, the examinee calls a phone number in California and is connected with a computer. The examinee enters the test paper number on the telephone keypad and then follows instructions. The first part of the test consists of reading from the test paper; thereafter, examinees repeat phrases and answer short questions. Finally, the examinee speaks freely about a given question. The test results are then retrieved from the company website.

The speech processing used by Ordinate Corporation was trained with the speech of native speakers and adapted for use by non-native speakers. It uses forced alignment to locate the relevant parts of the speech signal, a hidden Markov model-based speech recognizer, a pronunciation dictionary, and an expected-response network constructed from responses collected in thousands of administrations of the test (Ordinate, 1999: 2).

The PhonePass test result consists of five sub-scores on a scale from 2–8, which are combined to produce an overall score. The sub-scores include Listening Vocabulary, 30%; Repeat Accuracy, 30%; Pronunciation, 20%; Reading Fluency, 15%; and Repeat Fluency, 5%. In developing the test, a non-native norming group of 514 speakers of 40 foreign languages was formed. The mean overall score for this group was 5.2, with a standard deviation of 1.15. The standard error of the overall score is 0.2 (*ibid*: 2).

Repeated comparison of the results given by the PhonePass test and those obtained by human-rated measures of oral proficiency show that there is as much correlation between PhonePass scores and averaged human ratings as there is between one human rater and another (Bernstein, 1999: 8). The PhonePass test has been validated in relation to the spoken English tests given by a number of language institutes. The average correlation between PhonePass and these tests is 0.71, while the tests themselves had an average inter-rater correlation of .72 (*ibid*). It should be mentioned that this validation has been carried out on the overall score of the test, while the analysis presented here concerns the sub-score for pronunciation. An examination of some of the results of the PhonePass test is presented in Hincks (2001).

2.4.2 Other work using ASR for evaluation

With the aim of creating an automatic test for spoken Dutch, Cucchiarini *et al.* (2000) devised an extensive study that looked at the correlations between different aspects of human ratings of accented Dutch and machine scores provided by ASR. They found a high correlation between human ratings and machine-generated temporal measures such as rate of speech and total duration. In other words, speakers judged highly by the raters were also the faster speakers. However, the ASR did a poor job of assessing segmental quality. This was the aspect of speech that humans found to be most correlated with overall speech performance. There was thus a mismatch between what humans associated with good speech and what computers associated with good speech.

https://doi.org/10.1017/S0958344003000211 Published online by Cambridge University Press

3 Theoretical background

3.1 Theory implicit in design of product

Writers on the topic of CALL often decry the lack of theoretical motivation behind the design of CALL materials. This section examines how the underlying theories of two conflicting schools of thought have contributed to the design of the product tested in our study. In section 6, the product will be evaluated according to Chapelle's (2001) framework for software evaluation.

3.1.1 Talk to Me *and communicative language learning*

Software designers have been eager to incorporate speech recognition into CALL products because the ASR should allow application of communicative theories of language learning. This was clearly Auralog's point of departure in designing *Talk to Me*. From the beginning, the user is using his language to interact with the computer. The dialogue component is central to the software; the user chooses a response to an utterance and the dialogue proceeds according to the response. If the response is not understood by the computer, the user must repeat that response until he is understood; this is a sort of imitation of the negotiation of meaning (e.g. Pica, 1994) that takes place between human language speakers. Highlighting the 'worst' word of a response should aid in the noticing process necessary for learning. Providing multi-media-enriched sensory stimulation follows the recommendations of Krashen (1981) regarding the affective aspects of effective language learning. Finally, providing opportunity for simple practice of language should aid in the cognitive processes necessary for automating and learning (McLaughlin, 1990, as referenced in Levy, 1997).

3.1.2 Talk to Me *and audio-lingual language learning*

The communicative approach and SLA theories do not always have a lot to say about the acquisition of L2 phonology. The assumption has been that target-like pronunciation would follow as a result of language use in meaningful settings (Pennington & Richards, (1986:217). In order to provide specific pronunciation training, the program relies on a more outmoded school of thought, the audio-lingual theory of language teaching. If the dialogue sequences can be characterized as communicative, the exercises where phrases and words are repeated with visual feedback can be characterized as audio-lingual in that they rely on the imitation of models as opposed to meaningful communication. In these exercises the student drills words and phrases repeatedly, with the hope of getting a higher score from the program. Another type of drill associated with audio-lingualism is the presentation of minimal pair exercises to help in the discrimination of new sound contrasts. These would be simple to include in an ASR-based product, and were incorporated in the early products from Syracuse. Their exclusion from *Talk to Me* is perhaps unfortunate since minimal pair exercises have been demonstrated to benefit language learners (e.g. Callan , Tajima, Callan, Akahane-Yamada & Masaki, 2001); on the other hand, they are most useful when working between two specific L1s, while the Auralog product is designed for a worldwide market.

https://doi.org/10.1017/S0958344003000211 Published online by Cambridge University Press

3.2 Acquisition of L2 phonology

Commercial products are naturally inappropriate as vehicles for testing a specific theory of oral skills development. They are rather examples of 'prescriptive artefact design' (Jacobson 1994:143; as quoted by Levy 2002:71) where in this case a mixture of theories seem to be implicit in the product design. Our students were tested on their global pronunciation quality and this is what *Talk to Me* claimed to be able to assist. As advised by Felix (2002) we were using what was available rather than re-inventing the wheel. We were not, for example, testing whether a speaker of a specific L1 would be able to acquire a new sound contrast by using an ASR-based program, or whether it is more important to focus on prosodic or segmental training. Our results therefore have no broader bearing on the development of a unified theory of L2 phonological acquisition.

3.3 The tutor-tool distinction

The basic question of the research described in this paper was whether this product and its new approach to language training could deliver what it promised. What the product was promising, to an extent, was to duplicate human language teachers by means of automated responses. This is an example of using the computer as a tutor. Levy (1997) discusses the tutor-tool distinction as a useful conceptual framework for analysing the role of computers in language learning. Successful applications of the computer as a tool (e.g. word processing, e-mail) abound. Really successful applications of the computer as a tutor lie somewhere in the future. Programs such as *Talk to Me* are steps in that direction, in that they provide learner autonomy and flexibility (1997:205) and evaluate the learner (1997:180). However, the progress to a tutorial role is incomplete in that, as we have seen, the speech recognition cannot perform the crucial guiding role that a tutor needs to; furthermore, it is not always reliable in its feedback. Levy (1997:205) points out that a learner must be able to trust the computer, and goes so far as to say that "if the computer cannot be relied upon [as tutor], then it should probably not be allowed this relationship with the learner." In our study, however, students also met with a human tutor who helped them diagnose and correct their problems.

4 Test of effectiveness of *Talk To Me*

4.1 Framework

Our study involved two groups of students, a control group taking a course in the fall of 2000, and an 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 Immigrant Engineers offered at 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 pre- and post-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 so that they could use dictionaries more effectively. They did not receive their own pronunciation program. In

https://doi.org/10.1017/S0958344003000211 Published online by Cambridge University Press

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 otherwise generally the same for the two groups, but the teachers were different.

4.2 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), Armenian, Kurdish, Polish, Romanian, Russian and Tigrinya. Their average age was 42.

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

4.3 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. Each student also used the program for four hours in the company of a pronunciation tutor.

4.4 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. The 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.

5 Results

5.1 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.

https://doi.org/10.1017/S0958344003000211 Published online by Cambridge University Press

5.2 Overall scores

The mean overall score on the PhonePass test for the control group increased slightly from the pre-test to the post-test, from 4.48 to 4.74. The mean overall score for the experimental group increased by a smaller amount, from 4.4 to 4.51.

5.3 Pronunciation

5.3.1 Mean scores

Taken as wholes, neither the control nor the experimental group showed significant improvement in pronunciation as measured by the PhonePass sub-score for pronunciation.

Neither group produced changes in mean group scores from the pre-test to the post-test that exceeded 0.2, the amount Ordinate reports as significant. Figure 1 shows the pre- and post-testing results by student for the control group, and Figure 2 shows the same for the experimental group. The rightmost column in both figures represents the mean score for the groups. The mean score of the control group shows an insignificant decline from 4.7 to 4.6, and that of the experimental group no change at all at 4.1.

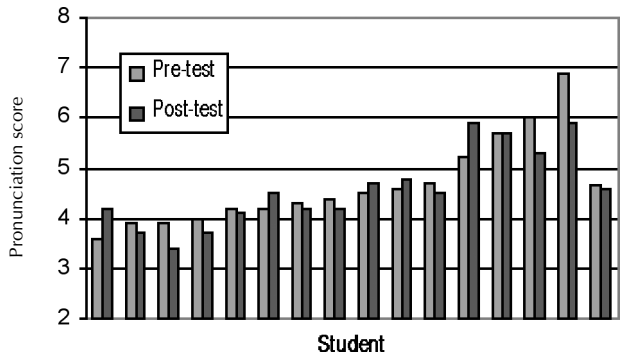


Fig. 1. Control group: Pre- and post-test scores in Pronunciation from PhonePass test. Mean group score in rightmost column.

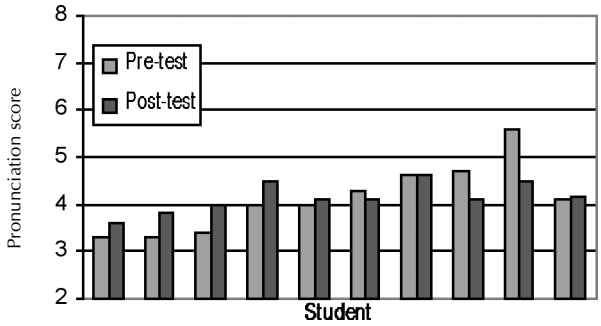


Fig. 2. Experimental group: Pre- and post-test scores in Pronunciation from PhonePass test. Mean group score in rightmost column.

https://doi.org/10.1017/S0958344003000211 Published online by Cambridge University Press

5.3.2 Individual scores

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 and 2. This subgroup showed measurable improvement in the experimental group, while there was little change 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.

5.3.3 Time on task

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.

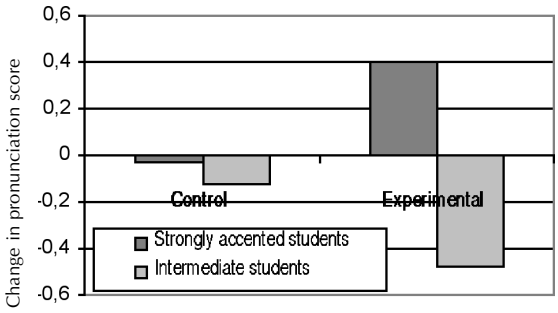


Fig. 3. Change in pronunciation score for students according to beginning level. Weak students improved significantly in the experimental group, while all other groups showed no improvement.

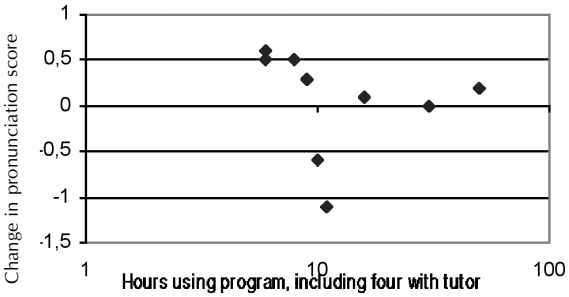


Fig. 4. Relationship between time spent using program and change in pronunciation score on PhonePass test.

https://doi.org/10.1017/S0958344003000211 Published online by Cambridge University Press

6 Discussion

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 (1984) 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.

6.1 Evaluating software

Chapelle (2001) offers a framework for evaluating CALL tasks. She suggests that a given piece of software be analysed according to the following six criteria: language learning potential, learner fit, meaning focus, impact, authenticity and practicality. In the following section, we discuss how *Talk to Me* meets these criteria.

6.1.1 Language learning potential

Chapelle defines language learning potential as "the extent to which an activity can be considered to be a language learning activity rather than simply an opportunity for language use" (2001:55). To do this, the task must provide a focus on form. *Talk to Me* does this by providing activities where the student can practise pronunciation at the phrase, word or phoneme level. The student can choose to practise all words starting with a particular consonant or containing a particular vowel, getting feedback in the form of a score. Instruction as to the articulation of the sounds is provided in the form of animations. Still, users can need guidance in finding the necessary information. 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. Having the student actively engaged in the diagnosis of his pronunciation difficulties is pedagogically desirable. The ideal automatic system would, however, both explain the error and point the student in the direction of the appropriate remedial activities.

Most SLA researchers consider that some form of feedback is necessary for language learning to take place. As discussed in section 2.1, visual feedback on prosody has been shown to help learners improve their intonation. However, all these positive studies concerned training conducted in conjunction with a teacher who could interpret the pitch curves and waveforms. *Talk to Me* provides a pitch curve and waveform to the student without the aid of expert interpretation. It is not certain that students are able to extract meaning from these signals (Neri, Cucchiarini & Strik, 2002). 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. Furthermore, when native speakers study the kind of scoring provided on their own pronunciation, they can find the scoring to be somewhat arbitrary.

https://doi.org/10.1017/S0958344003000211 Published online by Cambridge University Press

6.1.2 Learner fit

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 attempts 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 mimicking 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 their pronunciation is evidence that mimicry does not necessarily improve pronunciation. In fact, these results support the advice given by Morley (1991) that imitative speaking practice be used only for beginning students of a language.

6.1.3 Meaning focus

As defined by Chapelle, meaning focus is that "the learner's primary attention is directed toward the meaning of the language that is required to accomplish the task" (2001:56). This is certainly not taking place as the user drills the words and phrases of the pronunciation tasks in *Talk to Me*. In the dialogues, the user can choose a response; this injects a weak component of meaningfulness to the exercise.

6.1.4 Impact

A positive impact from a language learning activity helps the learners "develop their meta-cognitive strategies" (2001:57) so that learning can continue outside the classroom. Our students expressed high satisfaction with the program and were able to keep their own copies, providing them at least with the opportunity to continue working on their own. A follow-up study would need to be done to assess whether the students had developed a heightened awareness of the importance of correct articulation. A question here is whether the age of our students exposed them to the risks of fossilized pronunciation habits, making any advances in pronunciation extremely difficult.

6.1.5 Authenticity

The dialogues that a user has with *Talk to Me* are reasonably believable, dealing as they do with social or tourist situations. Speaking with a computer is, however, at present not terribly natural, though language technologists hope that it will become so in the future.

6.1.6 Practicality

Talk to Me is reasonably priced and runs on Windows 95 or better. Most of our students had no difficulty in either installing or using the program. It was clearly beneficial to

https://doi.org/10.1017/S0958344003000211 Published online by Cambridge University Press

also use the program in the company of a tutor, but students could continue work at home without problems.

6.2 Evaluating ASR for spoken language evaluation

6.2.1 Varieties of English: a problem for speech recognition

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. We had in fact ordered Am. E. programs, but the company had decided to sell only mixed-model programs. 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.

6.2.2 Student adaptation to negative feedback

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. In the post-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.* (2000) who found a negative correlation between practice with language learning software and speaking rate, indicating that increased attention to pronunciation could slow down speech. In a study comparing human judgements of spoken Dutch with automatic evaluation, Cucchiarini *et al.* (2000) 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.

7 Conclusions and future work

Extra pronunciation training using speech recognition-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 set of students did not improve if they were in the control group, despite the five hours of pronunciation tutoring they received. Though the small number of students makes it difficult to draw significant conclusions, the results indicate that ASR-based training could be useful for beginning students.

https://doi.org/10.1017/S0958344003000211 Published online by Cambridge University Press

This study is part of a larger project with the goal of creating oral skills software using an animated agent, a talking head, as an intelligent language tutor. Such a tutor would need to be endowed with sophisticated speech recognition in order to 'understand' students, and natural-sounding speech synthesis in order to communicate with them. As we have seen, speech recognition has a way to go before meeting these goals. However, it is beginning to play a role in language learning with some positive effects, and will hopefully continue to do so in the future.

Acknowledgements

My profound thanks to my colleagues at the Unit for Language and Communication at KTH for their support, to our LAN students for their participation and to my advisors David House and Björn Granström for their guidance.

References

Acton, W. (1984) Changing Fossilized Pronunciation. *TESOL Quarterly* **18**(1): 71–85.

Anderson-Hsieh, J., Johnson, R. and Koehler, K. (1992) The relationship between native speaker judgments of non-native pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning* **42**: 529–555.

Bannert, R. (1979) Rapport från uttalskliniken. *Praktisk Lingvistik* **1**, Lund University Department of Linguistics.

Bernstein, J. (1999) *PhonePass Data Analysis: Correspondence with Oral Interviews and First-Language Bias Analysis*. Ordinate Corp., Menlo Park, California.

Brown, G. and Yule, G. (1983) *Teaching the Spoken Language*. Cambridge: Cambridge University Press.

Bonaventura, P., Howarth, P. and Menzel, W. (2000) Phonetic annotation of a non-native speech corpus for application to computer-aided pronunciation teaching. *Proceedings InSTIL 2000*: pp. 10–17, University of Abertay, Dundee.

Callan D., Tajima, K., Callan, A., Akahane-Yamada, R. and Masaki, S. (2001) Neural Processes Underlying Perceptual Learning of a Difficult Second Language Phonetic Contrast. *Proceedings of Eurospeech 2001*(1): 145–148. University of Aalborg.

Chapelle, C. (2001) *Computer Applications in Second Language Acquisition*. Cambridge: Cambridge University Press

Chun, D. (1998) Signal analysis software for teaching discourse intonation. *Language Learning and Technology* **2**: 61–77.

Coniam, D. (1999) Voice recognition software accuracy with second language speakers of English. *System* **27**: 49–64.

Cucchiarini, C., Strik, H. and Boves, L. (2000) Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication* **30**: 109–119.

de Bot, K. and Mailfert, K. (1982) The Teaching of Intonation: Fundamental Research and Classroom Applications. *TESOL Quarterly* **16**(1): 71–77.

de Jong, J. and Bernstein, J. (2001) Relating PhonePass Overall Scores to the Council of Europe Framework Level Descriptors. *Proceedings of Eurospeech 2001*: 2803–2806.

Delmonte, R. (2000) SLIM prosodic automatic tools for self-learning instruction. *Speech Communication* **30**: 145–166.

https://doi.org/10.1017/S0958344003000211 Published online by Cambridge University Press

Derwing, T., Munro, M. and Carbonaro, M. (2000) Does popular speech recognition software work with ESL speech? *TESOL Quarterly* **34**(3): 592–603.

Derwing, T. and Rossiter, M. (2002) ESL Learners' perceptions of their pronunciation needs and strategies. *System* **30**: 155–166.

Egan, K. and LaRocca, S. (2000) Speech Recognition in Language Learning: A Must. *Proceedings of InStill 2000*: 4–7. University of Abertay, Dundee.

Eskenazi, M. (1999) Using automatic speech processing for foreign language pronunciation tutoring. *Language Learning and Technology* **2**(2): 62–76.

Felix, U. (2002) Teaching Languages Online: Deconstructing the Myths. Paper presented at *EuroCall 2002*.

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., Butzberger, J., Rossier, R. and Cesari, F. (2001) The SRI EduSpeak System: Recognition and Pronunciation Scoring for Language Learning. *Proceedings of InStill 2000*: 123–128. University of Abertay, Dundee

Hardison, D. (2002) Computer-assisted second-language learning: Generalization of prosody-focused training. *Proceedings of ICSLP 2002*: 1217–1220.

Hincks, R. (2001) Using speech recognition to evaluate skills in spoken English. Papers from *Fonetik 2001*: 58–61. Lund University Department of Linguistics.

Holland, M.V., Kaplan, J. and Sobol, M. (1999) A speech-interactive micro-world. *CALICO Journal* **16**(3): 339–359.

Jacobson, M. (1994) Issues in Hypertext and Hypermedia Research: Toward a Framework for Linking Theory-to-Design. *Journal of Educational Multimedia and Hypermedia* **3**(2): 141–154.

Jenkins, J. (2000) *The Phonology of English as an International Language*. Oxford: Oxford University Press.

Krashen, S. (1981) *Second Language Acquisition and Second Language Learning*. Oxford: Pergamon.

Levy, M. (2002) CALL by design: discourse, products and processes. *ReCALL* **14**(1): 58–84.

Levy, M. (1997) *Computer-Assisted Language Learning: Context and Conceptualization*. Oxford: Oxford University Press.

Macdonald, D., Yule, G. and Powers, M. (1994) Attempts to improve English L2 pronunciation: the variable effects of different types of instruction. *Language Learning* **44**: 75–100.

McLaughlin, B. (1990) Restructuring. *Applied Linguistics* **11**(2): 113–28.

Menzel, W., Herron, D., Bonaventura P. and Morton, R. (2000) Automatic detection and correction of non-native English pronunciation. *Proceedings of InSTIL 2000*, pp 49–56. University of Abertay Dundee.

Menzel, W., Herron, D., Morton, R., Pezzotta, D., Bonaventura P. and Howarth, P. (2001) Interactive Pronunciation Training. *ReCAll* **13**(1): 67–78.

Minematsu, N., Kurata, G. and Hirose, K. (2002) Corpus-based analysis of English spoken by Japanese students in view of the entire phonemic system of English. *Proceedings of ICSLP 2002*: 1213–1216.

Molholt, G. (1988) Computer-Assisted Instruction in Pronunciation for Chinese Speakers of American English. *TESOL Quarterly* **22**(1): 91–111.

Morley, J. (1991) The Pronunciation Component in Teaching English to Speakers of Other Languages. *TESOL Quarterly* **25**(3): 481–-520.

Neri, A., Cucchiarini, C. and Strik, H. (2002) Feedback in Computer Assisted Pronunciation Training: Technology Push or Demand Pull? *Proceedings of ICSLP 2002*: 1209–1212.

Ordinate Corporation (1999) *Validation summary for PhonePass SET-10: Spoken English Test-10, system revision 43*. Menlo Park, California.

Öster, A-M. (1999) Strategies and results from spoken L2 teaching with audio-visual feedback. TMH Quarterly Status and Progress Report 1-2 1999, Stockholm: KTH Department of Speech, Music and Hearing.

https://doi.org/10.1017/S0958344003000211 Published online by Cambridge University Press

Pennington, M. and Richards, J. (1986) Pronunciation Revisited. *TESOL Quarterly* **20**(2): 207–225.

Pica, T (1994) Research on Negotiation: What Does it Reveal About Second Language Learning Conditions, Processes and Outcomes? *Language Learning* **44**(3): 493–527.

Precoda, K., Halverson, C. and Franco, H. (2000) Effects of Speech Recognition-based Pronunciation Feedback of Second-Language Pronunciation Ability. *Proceedings of InSTIL 2000,* pp 102–105. University of Abertay Dundee.

Probst, K., Ke, Y. and Eskenazi, M. (2002) Enhancing foreign language tutors – In search of the Golden Speaker. *Speech Communication* **37**: 161–173.

Tomokiyo, L (2000) Handling Non-native Speech in LVCSR: A Preliminary Study. *Proceedings of InStill 2000*: 62–68, University of Abertay Dundee.

Townshend, B., Bernstein, J., Todic, O. and Warren, E. (1998) Estimation of spoken language proficiency. *Proceedings of ESCA Workshop on Speech Technology in Language Learning* (StiLL 98): 179–182. Stockholm: KTH Department of Speech, Music and Hearing.

Web sites

www.auralog.com
www.ordinate.com
www.eduspeak.com

https://doi.org/10.1017/S0958344003000211 Published online by Cambridge University Press