Computer-assisted Pronunciation Teaching based on Automatic Speech Recognition

Silke Witt Steve Young
Cambridge University Engineering Department
Trumpington Street
Cambridge CB2 1PZ
England

April 28, 1997

Abstract

Pronunciation teaching methods, as a part of computer assisted language learning systems, are currently limited in their ability to produce feedback on pronunciation quality. After an overview of previous work on pronunciation teaching, this article presents a pronunciation scoring algorithm based on automatic speech recognition, whereby scores at a phonemic level can be calculated. These "goodness of pronunciation" scores consist of a likelihood ratio between forced alignments and a maximum likelihood monophone loop. The results of evaluation experiments demonstrate the method's capability of detecting both individual mispronunciations as well as to give a general assessment of which sounds a student tends to pronounce badly.

1 Introduction

1.1 Aspects of Pronunciation Teaching

Before methods of computer-assisted pronunciation teaching can be devised, it is important to recognise the specific difficulties encountered in pronunciation teaching:

- Explicit pronunciation teaching requires the sole attention of the teacher to a single student; this poses a problem in a classroom environment;
- Learning pronunciation can involve a large amount of monotonous repetition, thus requiring a lot of patience and time from the teacher;
- With pronunciation being a psycho-motoric action, it is not only a mental task but also demands coordination and control over many muscles. Given the social implications of the act of speaking it can also mean that students are afraid to perform in the presence of others;
- In language tests the oral component is costly, time-consuming and subjective, therefore an automatic method of pronunciation assessment is highly desirable;

- Additionally, all arguments for the usefulness of computer-assisted language learning (CALL) systems apply here as well, such as being available at all times and being cheaper. See [Bailin 1995] for detailed reference.

All these reasons indicate that computer-based pronunciation teaching is not only desirable for self-study products but also for products which would complement the teaching aids available to a language teacher.

Having established the need for automatic pronunciation teaching tools, the next step is to determine which components of pronunciation to address. Roughly speaking, pronunciation quality is defined by its phonetic and prosodic features. For beginners, phonetic characteristics are of greater importance because these cause mispronunciations. With increasing fluency more emphasis should be on teaching prosody, i.e. intonation, stress and rhythm. This work however, is solely concerned with the assessment and correction of mispronunciations. This type of assessment can be done at the individual phoneme level as well as at the whole word or sentence level. The experiments presented here measure performance for assessing the quality of individual phonemes. Using scores at phoneme level makes it possible to give more detailed feedback of what has been spoken incorrectly in a word or sentence. In contrast, calculating a single score for a word or sentence provides little more information than whether the overall pronunciation of the utterance was native-like or whether is was rather poor. It is not possible to give more details about which mistakes have been made.

1.2 The Challenges of Computer-assisted Pronunciation Teaching

Although considerable research effort has been invested in the development of computer-assisted foreign language teaching systems, little attention has been paid to pronunciation teaching. This component of language teaching has the disadvantage that it is not possible to process and evaluate any oral response of a student using the standard means of interaction such as keyboard and mouse. Additionally, unlike with grammar or vocabulary exercises where there exist clearly defined wrong or right answers, in pronunciation exercises there exists no clearly right or wrong answer. A large number of different factors contribute to the overall pronunciation quality and these are also difficult to measure. Hence, the transition from poor to good pronunciation is a gradual one, and any assessment must also be presented on a graduated scale.

With the increasing performance of speech recognition over recent years, automatic pronunciation assessment is becoming feasible. However, existing speaker-independent recognition systems tend to perform badly when recognising non-native speech, [Gauvain et al. 1994]. Thus, applying speech recognition technology to the task of interactive language learning requires the introduction of new algorithms geared towards the specific requirements of non-native speech recognition.

1.3 State-of-the-art

Over the last decade several research groups have started to develop interactive language teaching systems incorporating pronunciation teaching based on speech recognition techniques. There was the "SPELL" project from [Hiller et al. 1993], which concentrated on teaching pronunciation of individual words or short phrases plus additional exercises for intonation, stress and rhythm. However, this system concentrated on one sound at a time, for instance the pair "thin-tin" is used to train the `th` sound, but it did not check whether the remaining phonemes in the word were pronounced correctly. Another early approach, based on dynamic programming and vector quantisation, by [Hamada et al. 1993], is likewise limited to word level comparisons between recordings of native and non-native utterances of a word. Therefore, their system required new recordings of native speech for each new word used in the teaching system. We will call this a text-dependent system in contrast to a text-independent one, where the teaching material can be adjusted without additional recordings. The systems described by [Bernstein et al. 1990] and [Neumeyer et al. 1996], were capable of scoring complete sentences but not smaller units of speech. The system used by [Rogers et al. 1994] was originally designed to improve the intelligibility of hearing or speech impaired people. It was text-dependent and evaluated isolated word pronunciations only. The system described by [Eskenazi 1996] was also text dependent and compared the log-likelihood scores produced by a speaker independent recogniser of native and non-native speech for a given sentence.

In contrast to these previous approaches, this work focuses on continuous speech at the level of phonemic transcriptions, i.e. producing pronunciation scores for each phoneme. The aims are to detect which phonemes were pronounced and to assess how close the pronunciation was to that of a native speaker. A future goal will be to correct the student with the help of various different forms of feedback, such as visualising the vocal tract, written explanation specifying how to pronounce certain sounds and much more.

2 Goodness of Pronunciation (GOP) Method

In order to design an algorithm for computing a "Goodness of Pronunciation" metric, the following assumptions were made:

1. The aim is to teach "comfortably intelligible" pronunciation;
2. All teaching is to be based on one single pronunciation set, in this case US American English, i.e. someone speaking British English would be corrected as well. Even though English has been used as the target language, this is only for evaluation purposes and the algorithm itself is generally applicable to any language pair;

3. It is assumed that the text spoken by the student is known at the moment of assessment, but need not be known when the models are trained. This is an important aspect for a flexible design of CALL systems.

Using a speech recogniser based on hidden Markov models (HMM), the output likelihoods computed during recognition for a phoneme contain information about how close the uttered phoneme was to the corresponding model. Under the simplifying but only approximately true assumption that an HMM phoneme model represents the "perfect" pronunciation, HMMs can be regarded as a stochastic model of pronunciation, so that low likelihood scores represent poor pronunciations and vice versa.

The goal is to have a score of correctness or confidence for each phoneme of a desired transcription for a sentence. This can be interpreted as the posterior probability that a speaker uttered phoneme p given the acoustics, O, and the set of all phonemes, Q:

$$GOP(p) \equiv log(P(p|O)) = log\left(\frac{P(O|p)P(p)}{\sum_{q\in Q} P(O|q)P(q)}\right). \tag{1}$$

Assuming that all phonemes are equally likely ($P(p) = P(q)$) and also that the sum in the denominator can be approximated by its maximum, the above equation rewrites as:

$$GOP(p) = log\left(\frac{P(O|p)}{\max_{q\in Q} P(O|q)}\right). \tag{2}$$

Such a score is based on keyword spotting and confidence measure techniques, see for instance [Knill and Young 1994] and [Young 1994].

This GOP score can therefore be found using two recognition passes of a sentence, the first uses forced alignment to transcriptions determined from a pronunciation dictionary, thus calculating $P(O|p)$. The second consists of a monophone loop permitting recognition of all possible sequences of phonemes, thus recognising the most likely phoneme sequence, i.e. $\max_{q\in Q} P(O|q)$:

$$GOP = S_{FA} - S_{PL}. \tag{3}$$

where S_{FA} is the Viterbi maximum log-likelihood score per frame for a phoneme segment of the forced alignment recognition pass,i.e.

$$S_{FA} \quad = \quad \frac{phoneme\ exit\ log\ prob - phoneme\ entry\ log\ prob}{number\ of\ frames\ of\ phoneme\ duration}. \tag{4}$$

Likewise, S_{PL}, the Viterbi maximum log-likelihood score per frame of a recognition pass using a phoneme loop, is calculated for the same speech segment. Because the phoneme loop segmentation can be different to that resulting from forced alignment, the average frame log-likelihood is calculated as the weighted sum of all phoneme segments occuring during the segment which is described by the number of frames it covers, i.e. $NF_{(PL,i)} = f_e - f_s$, f_e and f_s

being end and start frame number of each covered phoneme. The equation for S_{PL} thus takes the following form:

$$S_{PL} = \frac{\sum_{i=1}^{m} \left(S_{(PL,i)} NF_{(PL,i)}\right)}{number\ of\ frames\ of\ phoneme\ duration} \tag{5}$$

where $S_{(PL,i)}$ represents the maximum likelihood of the ith phoneme model in the desired speech segment. Describing this score verbally, one can say that this pronunciation score is defined as the ratio of the likelihood of the phoneme which should have been said (forced alignment) and the likelihood of the phoneme that actually has been said (phoneme loop).

3 Evaluation

In combination with a decision threshold, a GOP score as presented above can be used to measure the number of correctly pronounced phonemes in a given data set. A block-diagram of such a system is shown in Figure 1. The student's speech is digitised, then overlapping frames of 25 ms duration are converted every 10 ms into spectral feature vectors and used as input to the two recognition passes. The recogniser outputs are used to calculate the GOP score as defined in the previous section. The recognition components of the system are based on the HTK Toolkit, [Young et al. 1996]. Details concerning the models and spectral feature vectors used will be given in the next section. The final stage of error detection uses a threshold to decide whether a phoneme was correct or not, based on the GOP score in relation to a predetermined threshold.

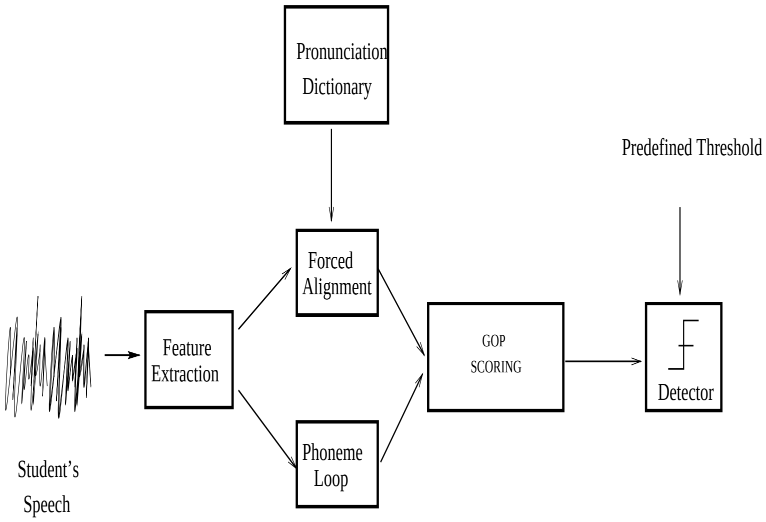

FIGURE 1 Block-diagram of a pronunciation scoring system

At the time of writing, fully annotated non-native speech was not yet

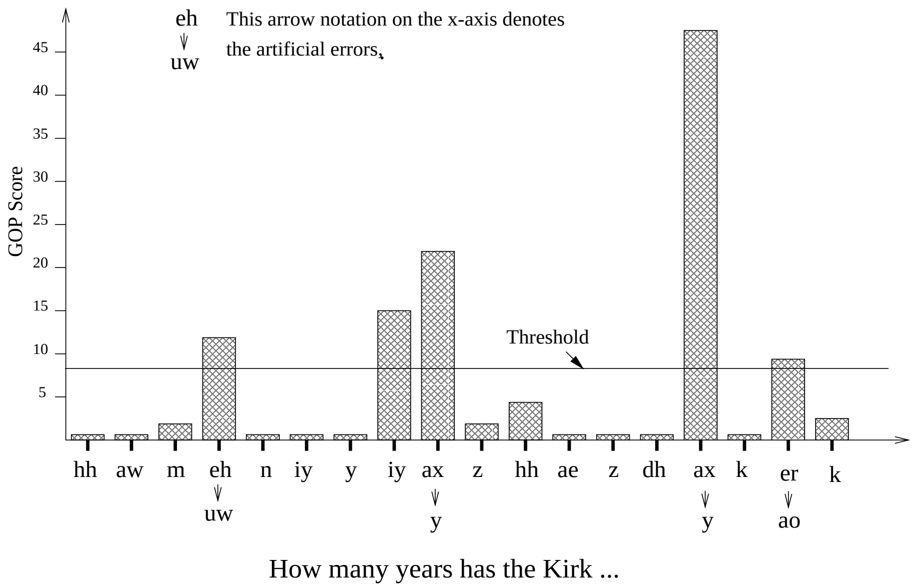

FIGURE 2 GOP scoring results for an example sentence

available. Hence, we produced artificial data to measure how a GOP score corresponds to a correctly spoken phoneme using the Resource Management Database. This database consists of continuous speech spoken from a variety of North American speakers. The dictionary used was based on a dictionary from Carnegie-Mellon University using a set of 48 phonemes for the phonetic transcriptions, including two silence models. From the RM data the artificial data were created by manipulating the pronunciation dictionary of the database so that the pronunciations were changed to contain different phonemes. For instance, all occurrences of the sound `aa` have been changed to `iy` and so forth. Thus, speech data with known locations of pronunciation errors was created.

To make a more detailed analysis of performance four decision types can be defined:

1. Correct Acceptance (CA): A phoneme has been pronounced correctly and was detected to be correct;
2. false Acceptance (FA) : A phoneme has been pronounced incorrectly and was detected to be correct;
3. Correct Rejection (CR): A phoneme has been pronounced incorrectly and was detected to be incorrect;
4. False Rejection (FR) : A phoneme has been pronounced correctly and was detected to be incorrect.

For a given threshold, statistics of all these four decision types can be collected. Defining scoring accuracy (SA) as $CA + CR$, SA can be plotted as a function of FA for a range of thresholds. These plots allow a system to be designed with optimal performance, i.e. optimal scoring accuracy for a given acceptance level of false acceptances.

An example of the results for a phrase containing artificially induced errors can be seen in Figure 2. By choosing a threshold of 8 in this example all induced errors will be detected, but also one correct phoneme `iy` will be rejected.

With this setup, the feedback to the student could be such that he or she is told which phonemes were pronounced incorrectly. Additionally, using the results of the phoneme loop, the student can be informed as to which phoneme has been said instead of the correct one. Again in this example, the phoneme loop classified the manipulated phonemes as `ih`, `er`, `ax` and `er` , whereas the phonemes actually spoken according to the original pronunciation dictionary were those in the upper line of the manipulated phonemes, i.e. `eh`, `ax`, `ax` and `er` in the order of occurrence. Thus, in the last two cases the de facto pronunciation was correctly recognised, in the first two instances closely related sounds were recognised. This correction feedback can be embellished in any desired way, such as including descriptions and pictures of the sound to be produced, playing example sounds and more.

4 Experimental Results

4.1 Optimisation of Model Type and Feature Vectors

Using the evaluation setup of the previous section several choices of HMM types and spectral feature vectors have been tested to determine which setup yields an optimal performance.

Firstly, we varied the type of hidden Markov models. The performance results of an experiment with speaker independent multiple (eight) mixture monophones, measuring scoring accuracy versus the false acceptance rate can be seen in Figure 3. For comparison, the performance using speaker independent tied-state triphones, which model more closely context and coarticulation is given as well. The poorer performance of triphone models is perhaps due to the fact that there are far fewer monophones (49) compared to the 112849 logical (6900 physical) triphones resulting in the monophones being more discriminative. Both model sets were trained on 73 speakers of the Resource Management Task, each speaker set consisting of 40 sentences.

In addition, to choosing a suitable model set, the selection of used feature vectors was also varied. In the previous experiments, a feature vector consisting of 13 Mel-frequency cepstral coefficients together with 13 delta, 13 acceleration and 3 energy coefficients (in short MFCC_D_A_E) was used. Next the model set was fixed to multiple mixture monophones and the choice of feature vectors has been varied. In Figure 4, results for Mel-frequency cepstral coefficients (MFCC) and Mel-frequency cepstral coefficients with delta and acceleration

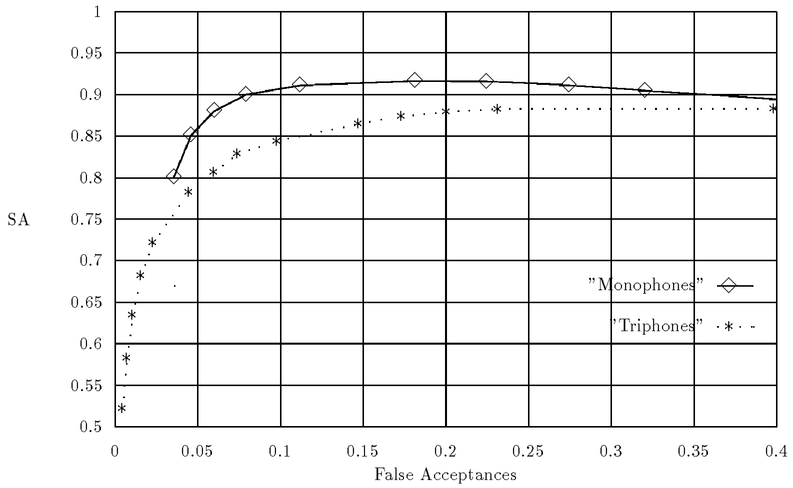

FIGURE 3 Scoring accuracy vs false acceptance for monophones and triphones

coefficients (MFCC_D_A) can be seen. The Mel-frequency cepstral coefficients together with delta, acceleration and energy coefficients perform best, which was expected as this feature set contains both static and dynamical information about the speech.

Summing up the results of Figure 3 and Figure 4, it can be deduced that the best system uses speaker independent multiple mixture monophones with Mel-frequency cepstral, delta and acceleration coefficients. With and a suitable threshold a scoring accuracy of 90% at a false acceptance rate of 8% can be achieved. These results show that — at least for this setup with artificially generated pronunciation errors — the GOP scoring method is a viable assessment tool.

4.2 Averaged Scores

Whereas so far all performance measurements have been for the individual score of each phoneme in a sentence, another experiment involved collecting average GOP scores for each phoneme over a large set of sentences for a single speaker. The scores of all occurrences of a given phoneme in a test text were accumulated and normalised by the number of occurrences. High scores indicate those phonemes, which are generally badly pronounced by a given speaker.

Typical errors for a German speaker are the pronunciation of the voiced and unvoiced `th`, the `r` and the `w` and `v`. The GOP scores for these sounds for both a native and a non-native student are shown in Figure 5. Generally, the score for the non-native speaker is much higher, thus this evaluation procedure can be used to broadly assess where the difficulties lie for a certain speaker. As a

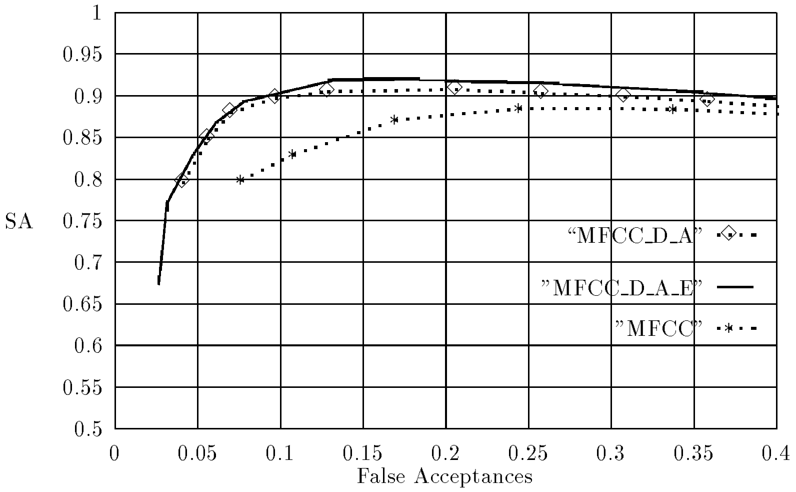

FIGURE 4 Scoring accuracy vs false acceptance for different feature. vector sets

contrast see Figure 6. Here the GOP score differences are shown for sounds which exists in both languages, and the differences shown are quite small.

5 Conclusions

5.1 Teaching System Design

For practical applications, any scoring method such as the one presented here will have to be embedded within an interactive language teaching system containing modules for error analysis, pronunciation lessons, feedback and assessment. These modules can take results from the core algorithm to give the student detailed feedback about the type of errors which occurred, using both visual and audio information. For instance, in those cases where a phoneme gets rejected because of too poor a score, the results of the phoneme loop indicate what has actually been recognised. This information can then be used for error correction. [Hiller et al. 1993] presented a useful paradigm for a CALL pronunciation teaching system called DELTA consisting of the four stages of learning:

1. Demonstrate the lesson audibly;
2. Evaluation Listening of the student ability with small tests;
3. Teach with pronunciation exercises;
4. Assess the progress made per lesson.

Other guidelines for such a design, findings from the areas of language learning [Kenworthy 1987], computer-assisted language learning [Bailin 1995] and from methods in the field of teaching speech and hearing impaired persons [Rogers et al. 1994] can be used. A different setup will be required for au-

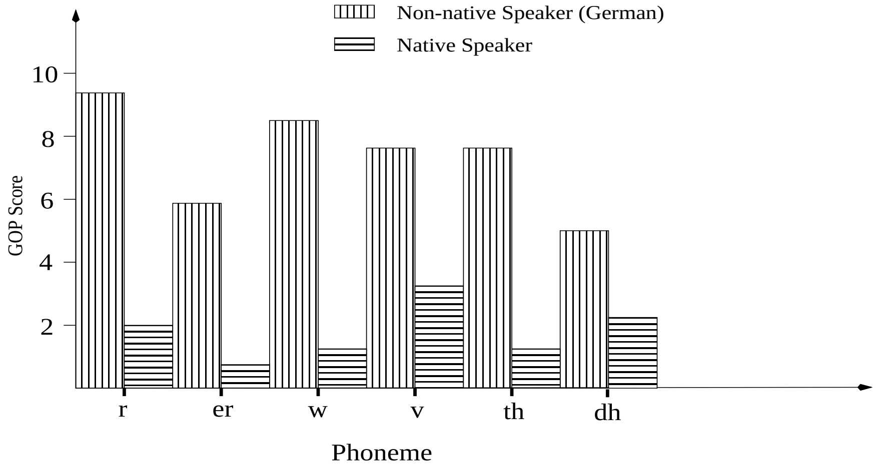

FIGURE 5 Native and non-native score comparison for typical error sounds

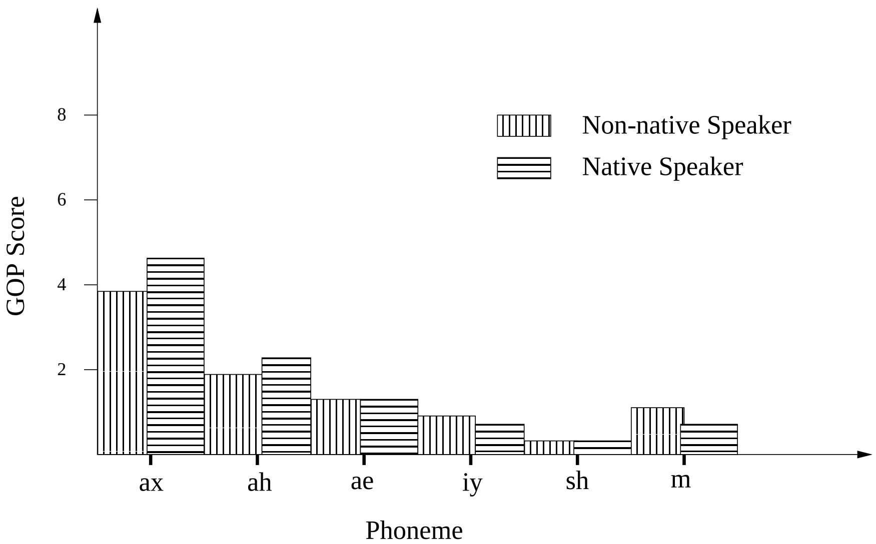

FIGURE 6 Native vs non-native comparison for sounds existing in both languages

tomated pronunciation assessment which will be based on the frame work of foreign language tests.

5.2 Conclusions and Future Work

The results of applying the algorithm introduced in this paper, which produces scores for each phoneme in an utterance, have shown that the automatic assessment of pronunciation is possible. The two different evaluation setups demonstrated the ability to evaluate individual sounds as well as providing an assessment of typical errors produced by a student.

One of the next steps will be to test the algorithm on non-native data, which has been hand-labelled. To this end a database of non-native speech has been recorded and is currently being annotated by trained linguists.

Altogether, the work presented here represents the beginning of a larger project in this area. Future work will include deriving methods to score at a word level and to use acoustic and duration modelling in order to increase the amount of information used in the scoring process. This is analogous to human judgement which is also based on knowledge about a large number of different features. Also, the algorithm does not yet include any further knowledge about the involved source and target language. With such information pronunciation networks could be built to model a range of common mistakes. Hence, there would be different networks for a German speaker learning English versus a Spanish speaker learning English. Finally, the work has to be extended to cover the dynamic components of pronunciation, such as rhythm and intonation.

References

Bailin, A. 1995. Intelligent Computer-Assisted Language Learning: A Bibliography. *Computers and the Humanities* 29:375–387.

Bernstein, J., M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub. 1990. Automatic Evaluation and Training in English Pronunication. In *ICSLP '90*, 1185–1188. Kobe, Japan.

Eskenazi, M. 1996. Detection of foreign speakers' pronunciation errors for second language training — preliminary results. In *ICSLP '96*. Philadelphia, PA, USA, Oct.

Flege, J.E., and K.L. Fletcher. 1992. Talker and listener effects on degree of perceived foreign accent. *J. Acoust. Soc. Am.* 91(1):370–389.

Gauvain, J.L., L.F. Lamel, G. Adda, and M. Adda-Decker. 1994. The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA Wall Street Journal Task. *ICASSP* I:557–564.

Hamada, H., S. Miki, and R. Nakatsu. 1993. Automatic Evaluation of English Pronunciation Based on Speech Recognition Techniques. *IEICE Trans. Inf. and Sys.* E76-D(3):352–359.

Hiller, S., E. Rooney, J. Laver, and M. Jack. 1993. SPELL: An automated system for computer-aided pronunciation teaching. *Speech Communication* 13:463–473.

Imai, T., A. Ando, and E. Miyasaka. 1995. A New Method for automatic generation of speaker dependent phonological rules. In *ICASSP*, 864–867. Detroit, USA.

Kenworthy, J. 1987. *Teaching English Pronunciation.* Longman.

Knill, K.M., and S.J. Young. 1994. Speaker Dependent Keyword Spotting for Accessing Stored Speech. Technical Report CUED/F-INFENG/TR 193. Cambridge, U.K.: Cambridge University, Oct.

Neumeyer, L., H. Franco, M. Weintraub, and P. Price. 1996. Pronunciation Scoring of Foreign Language Student Speech. In *ICSLP '96.* Philadelphia, PA, USA, Oct.

Rogers, C.L., J.M. Dalby, and G. DeVane. 1994. Intelligibility training for foreign-accented speech: A preliminary study. *J. Acoust. Soc. Am.* Vol. 96:no. 4, pt. 2.

Schmid, P., R. Cole, and M. Fanty. 1993. Automatically generated word pronunciations from phoneme classifier output. In *ICASSP*, 223–226. IEEE.

Young, S.J., J. Odell, D. Ollason, and P. Woodland. 1996. *The HTK Book.* Entropic Cambridge Research Laboratory.

Young, S.R. 1994. Recognition Confidence Measures: Detection of Misrecognitions and Out-of-vocabularyWords. Technical Report CMU-CS-94-157. Pittsburgh, PA 15213: Carnegie Mellon University, May.