AUTOMATIC PRONUNCIATION EVALUATION OF FOREIGN SPEAKERS USING UNKNOWN TEXT

N. Moustroufas and V. Digalakis

Dept. of Electronics and Computer Engineering
Technical University of Crete
Chania, 73100, GREECE

Corresponding Author:

Prof. Vassilis Digalakis,

Electronics and Computer Engineering Department,

Technical University of Crete, Chania, 73100, Greece.

Tel. +30-28210-37226, FAX +30-28210-37202,

Email: vas@telecom.tuc.gr

ABSTRACT

In this study we present various techniques to evaluate the pronunciation of students of a foreign language without any knowledge of the uttered text. Previous attempts have shown that it is feasible to evaluate the pronunciation of a non-native speaker by having implicit or explicit knowledge of the uttered text, provided that enough utterances are available. Our approach is to use characteristics of the mother tongue (SOURCE language) of the speaker in the evaluation of his/her pronunciation. We recorded 20 Greek students speaking English (TARGET language) and evaluated their pronunciation using algorithms that include characteristics of the SOURCE language (Greek). We show that the pronunciation scores that are based on both TARGET and SOURCE language characteristics have better correlation with the human scores than those based only on characteristics of the TARGET language. As in previous studies, we found that the best performing algorithms for automatic evaluation of pronunciation are based on speech recognition technology.

1 INTRODUCTION

In previous attempts (Bernstein, 1992; Neumeyer, Franco, Weintraub & Price, 1996;. and Franco, Neumeyer, Kim. & Ronen, 1997) on automatic evaluation of pronunciation there have been a number of scores upon which the evaluation was based and which seem to correlate quite well with the human ratings. Particularly in the studies of Neumeyer and Franco (Neumeyer et al, 1996; and Franco et al, 1997) we see that they used scores such as duration, syllabic-timing and hidden Markov model (HMM) log-likelihoods as valid indicators of pronunciation quality. Especially the log-posterior score in (Franco et al, 1997) produced the best correlation with the human scores. It was shown that the scores that were based on a number of utterances (speaker-level) had better performance, almost approaching human standards. On the other hand, the sentence-level correlations were low, indicating that further work had to be done. To overcome this effect in (Franco et al, 1997) some of the above scores were combined (log-posterior and duration scores) and that combination produced a 7% increase at the sentence-level correlations. However, sentence-level correlation of the machine and human scores was still low and this problem still remains unsolved.

The text uttered by the speaker in all the previous studies was in, one way or another, known to the system. For example in (Bernstein, 1992) the text was constant, i.e. all non-native speakers were recorded speaking over the same text, known to the system. Knowledge of the uttered text can lead to pronunciation scores that correlate very well with the human ratings. However, the algorithms were text-dependent, hence it was difficult to extend them to new texts. To overcome this effect, in the VILTS system (Neumeyer et al, 1996; and Franco et al, 1997) text-independent scoring algorithms were designed that were not based on the current text. However, even in the VILTS case, the uttered text should be available, so that the system could create time alignments and produce pronunciation scores. A number of interesting studies on pronunciation evaluation can be found in the special issue of Speech Communication (Neumeyer et al, 2000; Witt & Young, 2000; Cucchiarini et al, 2000; Franco et al, 2000; Kawai and Hirose, 2000; Delmonte, 2000; Ehsani et al, 2000; Yamada et al 2000).

This study extends previous work (Neumeyer et al, 1996; and Franco et al, 1997) by trying to evaluate the pronunciation of 20 Greek students of English without any knowledge of the text. To accomplish this, we introduce elements of the mother tongue (SOURCE language) of a non-native speaker. In our approach, we employ Gaussian Mixture Models (GMMs) and Gaussian Mixture

HMM-based speech recognizers to extract some characteristics of both English (TARGET language) and Greek (SOURCE language). We then combine the scores to produce a valid pronunciation score.

In our study we provide pronunciation scores at a sentence- or a speaker- (group of sentences) level. For feedback on pronunciation in computer-aided language learning (CALL) systems, scoring at a more detailed level is required, as has been done previously by Silke Witt, Goh Kawai and their colleagues (Kawai & Hirose, 1997; Kawai & Hirose, 1998; Witt & Young, 1997; Witt & Young, 1998a). Sentence- or section-level scoring is useful in language testing applications, as in the Autograder (Bernstein, 1992), VILTS (Neumeyer et al, 1996; and Franco et al, 1997) and the Ordinate PhonePass (Bernstein et al, 1998; and Ordinate, 2000) systems. One may argue that in language-testing systems the capability to evaluate pronunciation based on unknown text is of little use, since the text is usually known. This is true during the actual testing phase, when the text can be selected by the language-testing service providers, and hence the text is known by the language-testing system. Selecting a known (to the language-testing system) text during the testing may actually prevent the students from uttering a text that they have practiced very well, and thereby altering the meaning of the test. Nevertheless, pronunciation scoring based on unknown text allows the students to practice their overall pronunciation skills as much as they want using an automatic evaluation system, without requiring someone to enter the uttered text in the system. At the end, the official testing may be done using text that is known to the system. In any case and from an algorithmic point of view, the techniques that we present in this paper allow for a unified automatic pronunciation-scoring method that does not require knowledge of the text.

2 USING THE SOURCE LANGUAGE

We believe that the background of a speaker plays an important role when learning a foreign language. An automatic pronunciation-evaluation system should try to incorporate all available information of the speaker’s background into its scoring algorithms.

Information of the non-native mother tongue has been used in the past to assist automatic pronunciation-evaluation systems. In the Autograder system (Bernstein et al, 1990; Bernstein, 1992) a non-native database was used to calibrate the scoring algorithms using fixed text. Witt (Witt & Young, 1997) used non-native data to adapt the TARGET language recognizer in order to obtain better recognition performance for the non-natives. Witt later tried to improve the recognition performance on the non-natives (Witt & Young, 1998b) by using a linear model combination method on the SOURCE and TARGET acoustic models. However, the information of the SOURCE language recognizer was not used in the evaluation method. In (Kawai & Hirose, 1998), a bilingual recognizer was built using monophone models from the two languages that were operating in parallel. The phone networks for the various words of the TARGET language were built using knowledge of the phonotactic constraints and the common errors of the SOURCE-language speakers. The goal of this work was to identify errors at the phonetic level. In our work, instead of building a single recognizer, we use two parallel recognizers and combine their probability scores to evaluate the overall pronunciation quality of the non-native speaker, as we further explain in the remainder of the paper.

The automatic pronunciation-evaluation model (see Figure 1) used in previous approaches (e.g. Franco et al, 1997) includes a native speech corpus, which consists of sentences being uttered by native speakers of the TARGET language. This corpus is used to train an automatic speech recognizer for the TARGET language. There also exists a non-native speech corpus, which consists of speech from non-native speakers of the TARGET language on the text, which was used by the native corpus. The non-native speech corpus is evaluated by human raters and also by the machine using some form of machine score (e.g. log-posterior). Finally, correlation coefficients are computed for the human and machine scores.

(Figure 1 about here).

In this study the emphasis is placed upon the acoustic modeling of the languages under examination. When dealing with unknown text, we cannot use algorithms such as syllabic timing and phoneme duration, because we cannot be sure that the recognizer has identified the text correctly and has made the proper time alignments. Instead, we try to simulate the behavior of a human evaluator. We believe that a human evaluator, who is also a speaker of the SOURCE language, compares (mostly acoustic) characteristics of the TARGET and SOURCE languages and produces a result, which classifies a non-native speaker according to his/her pronunciation (see Figure 2). Hence, in our approach we use an additional native speech corpus, this time for the SOURCE language (Greek). Thus, we have two distinct native speech corpora. One is the native speech corpus of the TARGET language, which consists in our case of native speakers of English speaking English. The other native speech corpus consists of native speakers of Greek speaking Greek. The non-native speech corpus consists of 20 native speakers of Greek speaking English. We try to effectively combine the results returned from the two recognizers (TARGET & SOURCE) to produce a valid pronunciation score, which does not depend upon the text being uttered by the non-native speakers.

(Figure 2 about here).

3 THE DATABASE

3.1 Non-Native Speech Corpus

To meet the requirements of the present study we had to develop a small database of non-native speech. Speech was recorded from 20 natives of Greek speaking English (non-native corpus). The recordings took place in a quiet office using a high-quality dynamic microphone. The recordings were performed using the same data-collection procedures that were used for the creation of widely available corpora like the WSJ (Doddington 1992). The text upon which the non-natives were recorded comprised of 114 sentences of varying difficulty. The collected data consisted of read speech, and the text included abstracts of live interviews and sections of articles. The text also included interrogative and imperative speech. The total number of utterances recorded was 20 students reading 114 sentences each, for a total of 2280 utterances.

We have to mention that at least 4 speakers of native level were included in the non-native corpus. In general, the level of pronunciation of most of the non-native speakers was average.

3.2 Native Speech Corpus

We used two distinct native speech corpora. To model the natives of English we used the Wall Street Journal corpus (WSJ), which consists of recorded speech of Americans upon articles of the Wall Street Journal newspaper. To model the natives of Greek, we used the Greek speech corpus of Logotypografia (ENET) (Digalakis et al 2003), which was developed at the Technical University of Crete (T.U.C) and consists of recorded speech of Greeks upon articles of the Eleftherotypia newspaper.

For the GMM approach, we used 40 speakers from each class (Greek males, Greek females, American males, American females) with 75 utterances each (total 12000 utterances). We created 4 GMMs (one for each class) and used the log-likelihoods as machine scores.

For the HMM-based speech recognizers we incorporated 4 gender-specific, large-vocabulary speech recognition systems that were based on genonic HMMs (Digalakis & Murveit, 1994), one for each class as mentioned earlier, from the same speech databases (WSJ and ENET). We used trigram language models in our experiments.

For both the GMM and the HMM models, we used a six-feature front end that outputs 12 cepstral coefficients, cepstral energy, and their first- and second-order differences. The cepstral features are computed from an FFT filterbank.

3.3 Human Scoring

We used 3 expert human raters from the Language Research and Resource Center (L.R.R.C) of the Technical University of Crete (T.U.C.) to rate the non-native speech corpus. The rating was done on a 1 to 5 scale. Criteria were intonation, phoneme pronunciation and expression of the proper meaning of the sentence. The main focus was on pronunciation and not on individual or local accent. A non-native speaker would get a 5 if he had very good intonation, clear and proper pronunciation of individual phonemes and was expressing the right meaning of the sentence, no matter if he had a foreign accent.

The human scores are the reference against which the performance of the automatic scoring systems should be tested and calibrated; as such, it is important to assess the consistency of these scores between the raters themselves. To measure human consistency, we use simple linear correlation techniques (Montgomery & Runger, 1999). In order to measure the correlation between the raters we had a randomly-selected subset (~20%) of the non-native corpus rated by all raters, which we shall call the 20%-overlap. Two types of correlation were computed; at the sentence level pairs of corresponding ratings for all the individual sentences were correlated. At the speaker level, first, the scores for all the sentences from each speaker were averaged, and then the sequence of pairs of corresponding average scores for each of the speakers was correlated. In Table I, we show the results for sentence/speaker level correlation between each rater.

(Table I about here)

The level of correlation is reasonably uniform across all raters with the correlation between rater 1 and 2 being slightly lower. The correlations at the speaker level are consistently higher than those at sentence level, reflecting that the average scores based on several sentences are more reliable than those based on single sentences. The average correlation between raters at the sentence level is 0.61 while at speaker level it is 0.75. We have also computed what is referred to as “open correlation”, that is, the correlation between a rater and the average of all the others (Table II). We have also computed the average correlation at sentence and speaker level for all open correlations, which was found to be 0.68 at sentence level and 0.80 at speaker level. These values suggest an upper bound on the level of correlation we expect the automatic evaluation system to have with the human scores.

(Table II about here)

Descriptive statistics were obtained over the whole set of 2,280 human scores of non-native data of 20 speakers. The histogram of the scores, using a scale of 1 to 5 described earlier, from all raters for all sentences is shown in Table III. For the 20%-overlap subset, since the sentences were selected randomly we had similar statistics.

(Table III about here)

The scores have a rather symmetrical form around score 3, which reflects the fact that the non-native speakers' pronunciation was at average level. However, we see a rather high percentage (23%) of 5 scores, which reflects the fact that the non-native corpus included at least 4 speakers (20%) that had native-level pronunciation. In Table IV, the mean and standard deviation of the scores given by each rater are shown. The means differ at most by a half point (in rater 3), and the standard deviations are reasonably similar. The means and standard deviations of the scores on the 20%-overlap were very similar to the corresponding statistics computed over the whole set of sentences.

(Table IV about here)

We asked the raters to evaluate each sentence according to its "difficulty". The difficulty was determined by the rater based on the number of less frequent (uncommon) English words that it contained and the overall pronunciation difficulty. Each rater gave her own difficulty score for each sentence for all sentences contained in the 20%-overlap subset. We then averaged the difficulty scores from all raters for each sentence and computed the average of the pronunciation scores of the raters for each sentence difficulty. The results are presented in Table V.

(Table V about here)

Difficulty scores 1 and 5 were eliminated by the effect of averaging, thus in Table V there are no scores in these positions. However we see that as the difficulty level increases, the average pronunciation score lowers, which is quite reasonable.

4 AUTOMATIC PRONUNCIATION SCORING

In our attempt to evaluate the pronunciation of foreign students without any knowledge of the text uttered, we incorporated 2 different methods: the GMMs and HMM-based speech recognizers.

4.1 Gaussian Mixture Models

In this approach we use the GMM *log-likelihood* as a pronunciation score. The underlying assumption is that the logarithm of the likelihood of the speech data using GMMs obtained from native speakers is a good measure of the similarity between the native speech and the student speech. If $O_{NN} = \{o_1, o_2, .., o_T\}$ is the sentence uttered by a non-native speaker, then the total log-likelihood for this sentence is

$$L(O_{NN} \mid \Phi_{TARGET}) = \log f(O_{NN} \mid \Phi_{TARGET}) = \sum_{k=1}^{T} \log f(o_k \mid \Phi_{TARGET}),$$

where $L(O_{NN} \mid \Phi_{TARGET})$ represents the total log-likelihood of a non-native sentence O_{NN} computed with the GMMs that were obtained from the native speech data of the TARGET language. To normalize for the effect of the sentence length, we divide by T, which is the total number of frames in a sentence. Finally, we define a pronunciation score for a sentence O_{NN} based on the TARGET language as:

$$L = \frac{L(O_{NN} \mid \Phi_{TARGET})}{T} \tag{1}$$

In order to use the SOURCE language we proceed in the following way:

- If $L(O_{NN} \mid \Phi_{TARGET}) > L(O_{NN} \mid \Phi_{SOURCE})$, then the sentence O_{NN} is pronounced well by the speaker, else
- If $L(O_{NN} \mid \Phi_{TARGET}) < L(O_{NN} \mid \Phi_{SOURCE})$, then the sentence O_{NN} is NOT pronounced well by the speaker.

Following this line of reasoning, we can use a linear or non-linear combination of $L(O_{NN} \mid \Phi_{SOURCE})$ and $L(O_{NN} \mid \Phi_{TARGET})$ as a pronunciation score. One simple solution to this problem is to use the difference defined as:

$$LDIFF = \frac{L(O_{NN} \mid \Phi_{TARGET})}{T} - \frac{L(O_{NN} \mid \Phi_{SOURCE})}{T}, \qquad (2)$$

where $L(O_{NN} \mid \Phi_{SOURCE})$ stands for the total log-likelihood of a non-native sentence O_{NN} computed with the GMMs that were obtained from the native speech data of the SOURCE language. Since $L(O_{NN} \mid \Phi_{SOURCE})$ and $L(O_{NN} \mid \Phi_{TARGET})$ represent log-likelihood scores of the SOURCE and TARGET GMMs, respectively, their difference is equivalent to the logarithm of the likelihood ratio between the two GMMs. The log likelihood-ratio of two models is used extensively in decision and detection problems and this was our motivation for the selection of this score. In addition, we normalize the log likelihood ratio by the length of the sentence, *T*, to compensate for the variable length of the spoken utterances.

4.2 Hidden Markov Models

In the HMM approach we use the *confidence score* as a valid pronunciation predictor. The confidence for a sentence O_{NN} can be defined as (e.g. Witt and Young 1997):

$$C(O_{NN}) \approx \log p(\overline{W} \mid O_{NN}) = \log \frac{p(O_{NN} \mid \overline{W}) p(\overline{W})}{\sum_{W} p(O_{NN} \mid W) p(W)}.$$

This equation states that the confidence is analogous to the *a posteriori* probability of the word sequence $\overline{W}$ recognized by the recognizer, given the actual utterance data O_{NN}. The confidence score defined above is directly analogous to the log-posterior score described in (Franco, Neumeyer, Kim and Ronen, 1997), which was shown to be a very good predictor of the pronunciation quality. If we can put more emphasis on the acoustic model over the language model, then the confidence score can be a good pronunciation predictor. To control the relative weights of the acoustic and linguistic scores during recognition, we used a Grammar Probability Weight

(GPW). Lower values of GPW give a higher weight on the HMM observation probabilities and a lower weight on the language-model probability and, therefore, the confidence computation is influenced mainly by the acoustic model. A pronunciation score for a sentence O_{NN} based on the TARGET language can be defined as:

$$C = C(O_{NN} \mid \lambda_{TARGET}) \ , \tag{3}$$

where $C(O_{NN} \mid \lambda_{TARGET})$ represents the confidence score computed for a sentence O_{NN} using the HMMs obtained from the native speech data of the TARGET language.

Following the line of reasoning that we analyzed in the GMM section, we can use different combinations of TARGET and SOURCE language confidence scores for a given sentence O_{NN}, as a pronunciation score for the sentence. In the present study we used two combinations:

- The Difference

$$CDIFF = C(O_{NN} \mid \lambda_{TARGET}) - C(O_{NN} \mid \lambda_{SOURCE}) \tag{4}$$

- The Normalized Confidence

$$CNORM = \frac{C(O_{NN} \mid \lambda_{TARGET})}{C(O_{NN} \mid \lambda_{TARGET}) + C(O_{NN} \mid \lambda_{SOURCE})} \ , \tag{5}$$

where $C(O_{NN} \mid \lambda_{SOURCE})$ stands for the confidence score computed for a sentence O_{NN} using the HMMs obtained from the native speech data of the SOURCE language. Since we compute the confidence using log posterior probabilities, the score CDIFF corresponds to a log-likelihood ratio of the TARGET and SOURCE language recognizers. The score CNORM can be rewritten as:

$$CNORM = \frac{C(O_{NN} \mid \lambda_{TARGET})}{C(O_{NN} \mid \lambda_{TARGET}) + C(O_{NN} \mid \lambda_{SOURCE})} = \frac{C(O_{NN} \mid \lambda_{TARGET}) / C(O_{NN} \mid \lambda_{SOURCE})}{1 + C(O_{NN} \mid \lambda_{TARGET}) / C(O_{NN} \mid \lambda_{SOURCE})} \tag{6}$$

The justification for this score is that, as the confidence of the TARGET recognizer becomes much larger than the SOURCE recognizer, the CNORM score converges to 1, whereas when the confidence of the TARGET recognizer is much smaller than the SOURCE recognizer the CNORM score approaches zero.

5 EXPERIMENTAL RESULTS

To evaluate the pronunciation scoring algorithms we used 114 sentences from 20 Greek speakers with various levels of proficiency in English.

5.1 GMMs

We have experimented using different number of Gaussians to implement the GMMs. The results are shown in Table VI.

(Table VI about here)

We note that the LDIFF score, which is the difference of TARGET and SOURCE log-likelihoods, outperforms the L score, the log-likelihood of the TARGET language alone. We also note that, as the number of Gaussians increases, there is also an increase in correlation in all levels and all types of scores. This should be expected; as the large number of utterances (12000) used to train the GMMs are adequate to estimate a larger number of Gaussians that model the acoustic space better. We see that the L score, although having poorer correlation with human scores than LDIFF, has a larger increase in correlation as the number of Gaussians increases. On the other hand, we must note that the LDIFF score outperforms L even when we use 50 Gaussians for the former and 1024 Gaussians for the latter.

We have also evaluated the correlation as a function of the amount of sentences per speaker, for both score types. The results are presented in Figure 3 for the best-performing case of 1024 Gaussians.

(Figure 3 about here)

In all cases the LDIFF has better performance over the L score. The L score correlation level reaches at maximum a value of 0.2, when the LDIFF correlation level reaches 0.6 for 1024 Gaussians.

5.2 HMMs

We have experimented using different values of the GPW parameter and the results are shown in Table VII.

(Table VII about here)

We see that sentence level correlation, for all values of GPW, is higher when using either CDIFF or CNORM score over C. Particularly CNORM produces the best sentence level correlation of 0.4913 when GPW=0.01. At the speaker level, we note that the CDIFF and CNORM scores have better performance over the C confidence score, which is based only on the TARGET language. Moreover, when using low GPW values like 0.01, where emphasis is placed mostly on acoustic scores, CDIFF and CNORM produce the best results. Especially CNORM produces a 10.9% increase in sentence level (from 0.4432 to 0.4913) and 2.7% increase in speaker level (from 0.7753 to 0.7966) over the C score.

We have also evaluated the correlation as a function of the amount of sentences per speaker, for these score types. The results are presented in Figures 4 through 7.

(Figures 4-7 about here)

The figures for the correlations are similar for the different values of GPW. We do note, however, that the different values of the GPW have an effect on the convergence rate. For example, for 2 sentences, the speaker-level correlations of the CNORM score are 0.6334, 0.6753, 0.5592 and 0.6334, for GPWs of 0.001, 0.01, 0.1 and 1, respectively. Hence, a value of 0.01 for the GPW achieves a faster convergence to the final speaker-level correlation. We can also see that the CNORM score has a very good performance almost in all cases. CNORM and CDIFF perform quite well, especially for a small amount of sentences (1-10). CDIFF has a good performance for greater values of GPW, like 0.1 or 1. We must note that when we use 10 sentences the speaker-level correlation of the CNORM score for a GPW of 0.01 is 0.7581, which corresponds to 95% of its maximum value (0.7966) calculated using 114 sentences. In the case of the CNORM score the correlations are very close to the human standard level of 0.8.

6 SUMMARY

We have presented a number of algorithms for automatic evaluation of pronunciation of foreign student speech when the text uttered is unknown. We designed algorithms that include characteristics of the mother tongue (SOURCE language) of the speaker and we showed that these algorithms had better performance over similar algorithms that depend upon the TARGET language only. We showed that using these algorithms we had a 10.9% increase in sentence level correlation and 2.7% increase in speaker level correlation. When using 10 sentences from a speaker we can make a valid prediction of his/her pronunciation according to human standards. However, sentence level correlation is still low and this fact indicates that there is a lot of work to be done to be able to predict the pronunciation of a speaker using a single utterance.

7 REFERENCES

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D. and Weintraub, M. (1990). “Automatic Evaluation and Training in English Pronunciation”. *In Proceedings of ICSLP 1990,* Kobe, Japan, 1990.

Bernstein, J. (1992). “Automatic grading of English spoken by Japanese students,” *SRI International Reports*, Project 2417, 1992.

Bernstein, J., De Jong, J.H.A.L., Pisoni, D., & Townshend, B. (2000). “Two Experiments on Automatic Scoring of Spoken Language Proficiency”. *In Proceedings of InSTIL2000: Integrating Speech Technology in Learning (pp. 57-61), University of Abertay Dundee,* Scotland, August, 2000.

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.

Delmonte, R. (2000). “SLIM Prosodic Automatic Tools For Self-Learning Instruction.” *Speech Communication,* **30**, 145-166.

Digalakis V. and Murveit, H. (1994). “GENONES: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer.” *In Proceedings of ICASSP94,* Sydney, 1994.

Digalakis, V., Oikonomidis, D., Pratsolis, D., Tsourakis, N., Vosnidis, C., Chatzichrisafis N. and Diakoloukas, V. (2003). “Large Vocabulary Continuous Speech Recognition in Greek: Corpus and an Automatic Dictation System”. *In Proceedings of Eurospeech’03*, September 2003.

Doddington, G. (1992). “CSR Corpus Development.” *In Proceedings of the ARPA Workshop on Spoken Language Technology,* February 1992.

Ehsani, F., Bernstein, J., and Najmi, A., (2000). “An Interactive Dialog System For Learning Japanese.” *Speech Communication,* **30**, 167-177.

Franco, H., Neumeyer, L., Kim, Y. & Ronen, O. (1997). “Automatic Pronunciation Scoring for Language Instruction.” *In ICASSP ’97*, Munich, Germany, April 1997.

Franco, H., Neumeyer, L., Digalakis, V. and Ronen, O. (2000). “Combination Of Machine Scores For Automatic Grading Of Pronunciation Quality.” *Speech Communication,* **30**, 121-130.

Kawai, G. and Hirose, K. (1997). “A CALL System Using Speech Recognition to Train the Pronunciation of Japanese Long Vowels, the Mora Nasal and Mora Obstruents.” *In Proceedings of EuroSpeech ’97*, Rhodes, Greece, September 1997.

Kawai, G. and Hirose, K. (1998). “A Method for Measuring the Intelligibility and Non-nativeness of Phone Quality in Foreign Language Pronunciation Training.” *In Proceedings of ICSLP’98, Sydney, Australia, November-December 1998.*

Kawai, G. and Hirose, K. (2000). “Teaching The Pronunciation Of Japanese Double-Mora Phonemes Using Speech Recognition Technology.” *Speech Communication,* **30**, 131-143.

Montgomery, D., C. & Runger, G., C. (1999). “Applied Statistics and Probability for Engineers,” 2nd ed, John Wiley and Sons, New York.

Neumeyer, L., Franco, H., Weintraub, M. & Price, P. (1996). “Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech.” *In Proceedings of ICSLP ’96*, Philadelphia, PA, USA, October1996.

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

Ordinate (2000). “Validation summary for PhonePass SET-10: Spoken English Test-10, System Revision 43”. Menlo Park, CA.

Witt, S. & Young, S. (1997). “Language Learning Based on Non-Native Speech Recognition.” *In Proceedings of EuroSpeech ’97*, Rhodes, Greece, September 1997.

Witt, S. & Young, S. (1998a). “Performance Measures for Phone-Level Pronunciation Teaching in CALL.” *In Proceedings of InSTIL1998: Integrating Speech Technology in Learning,* Marholmen, Sweden, 1998.

Witt, S. & Young, S. (1998b). “Estimation of Models for Non-native Speech in Computer-assisted Language Learning Based on Linear Model Combination.” *In Proceedings of ICSLP’98, Sydney, Australia, November-December 1998.*

Witt, S. & Young, S. (2000). “Phone-Level Pronunciation Scoring And Assessment For Interactive Language Learning.” *Speech Communication,* **30**, 95-108.

Yamada, Y., Javkin, H., and Youdelman, K. (2000). “Assistive Speech Technology For Persons With Speech Impairments.” *Speech Communication,* **30**, 179-187.

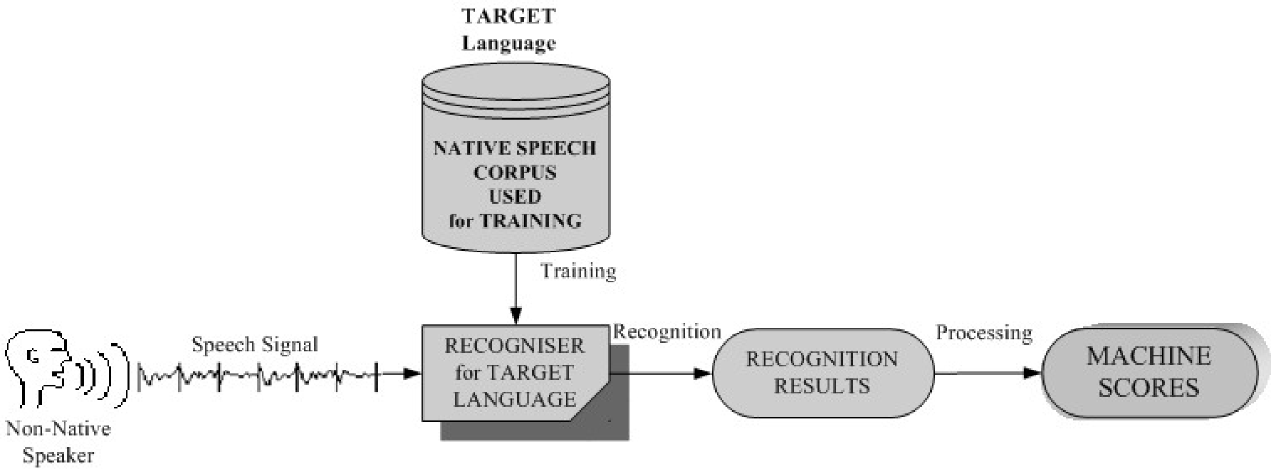

Figure 1: The VILTS system. A non-native speaker is evaluated on his pronunciation of the TARGET language according to how close his speech characteristics are to those of the native speakers. The text must be known to the system.

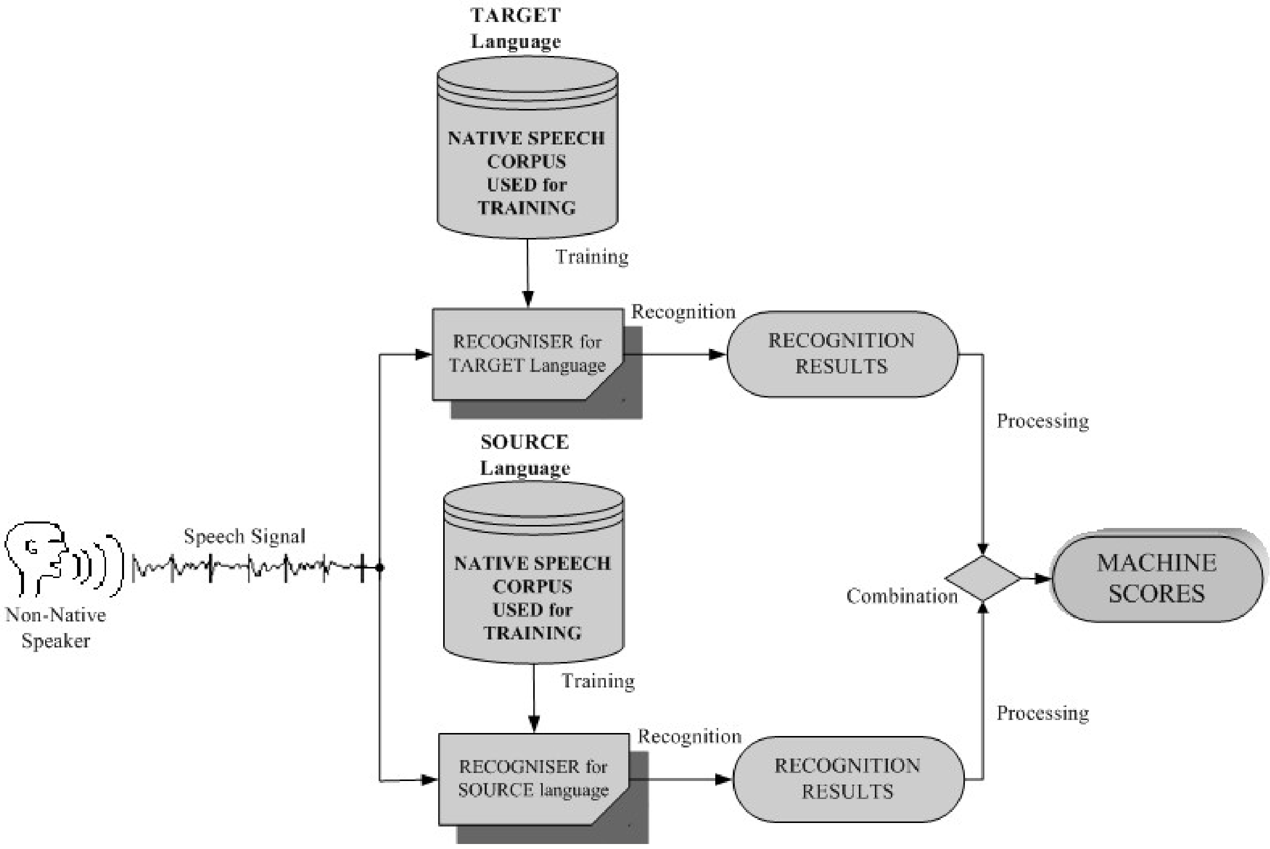

Figure 2: Present study. A non-native speaker is evaluated on his pronunciation of the TARGET language according to how close his speech characteristics are to those of the native speakers of the TARGET language AND how far are his speech characteristics from the native speakers of the SOURCE language (mother tongue). The text uttered is unknown to the system.

Rater Id	1	2	3
1	1.00/1.00	0.54/0.69	0.61/0.76
2		1.00/1.00	0.67/0.80
3			1.00/1.00

Table I: Sentence and Speaker-level correlations between raters.

Level	Rater Ids			Average
	1	2	3	
Sentence	0.63	0.67	0.73	0.68
Speaker	0.76	0.80	0.84	0.80

Table II: Sentence and Speaker-level open correlations between raters. Open correlation for a rater is computed against the average of the other raters.

Score	1	2	3	4	5
Percentage (%)	6	17	29	25	23

Table III: Histogram of scores across all raters for all the scores.

Rater	1	2	3	Average
Mean	3.6	3.6	3.1	3.4
St. Dev	1.2	1.0	1.3	1.2

Table IV: Means and standard deviations of scores from each rater.

Average Sentence Difficulty	Mean Score
1	-
2	3.53
3	3.40
4	3.24
5	-

Table V: Means of scores for each sentence difficulty. The sentence difficulty is the average difficulty from all raters.

Number of Gaussians	Machine Score	Correlation Coefficient	
		Sentence Level	Speaker Level
50	LDIFF	0.2295	0.3434
	L	-0.0136	0.0164
128	LDIFF	0.3385	0.4693
	L	0.0298	0.0809
256	LDIFF	0.3719	0.5178
	L	0.0532	0.1147
512	LDIFF	0.4096	0.5554
	L	0.0760	0.1489
1024	LDIFF	0.4286	0.5694
	L	0.0953	0.1731

Table VI: Sentence and speaker level correlations between human and machine scores in the case of the GMMs for different number of Gaussians. We use 20 non-native speakers and 114 utterances per speaker.

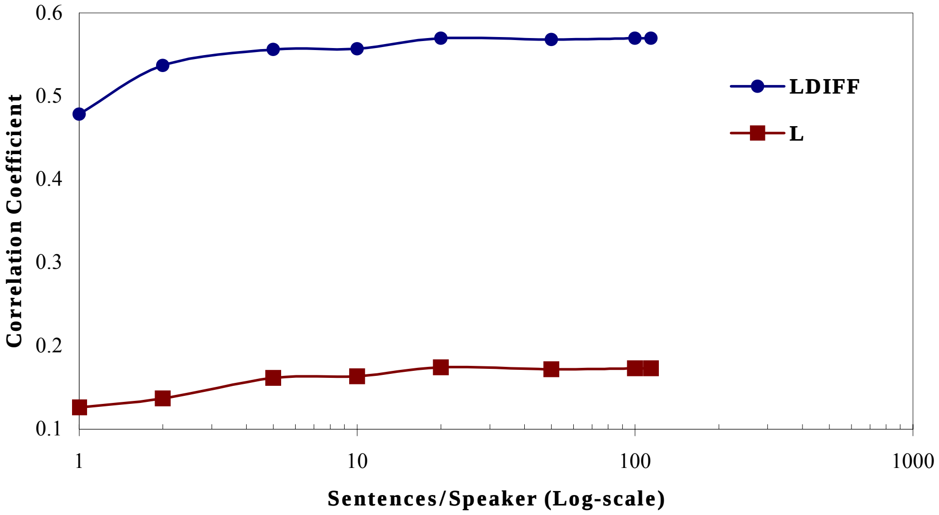

Figure 3: Correlation coefficient as a function of the number of sentences for GMMs with 1024 Gaussians

GPW	Machine Score	Correlation Coefficient	
		Sentence Level	Speaker Level
0.001	C	0.4426	0.7748
	CDIFF	0.4862	0.7826
	CNORM	0.4909	0.7967
0.01	C	0.4432	0.7753
	CDIFF	0.4867	0.7824
	CNORM	0.4913	0.7966
0.1	C	0.4395	0.7706
	CDIFF	0.4829	0.7752
	CNORM	0.4871	0.7894
1	C	0.4309	0.7511
	CDIFF	0.4796	0.7813
	CNORM	0.4692	0.7953

Table VII: Sentence and speaker level correlations between human and machine scores in the case of the HMMs for different values of the GPW. We use 20 non-native speakers and 114 utterances per speaker.

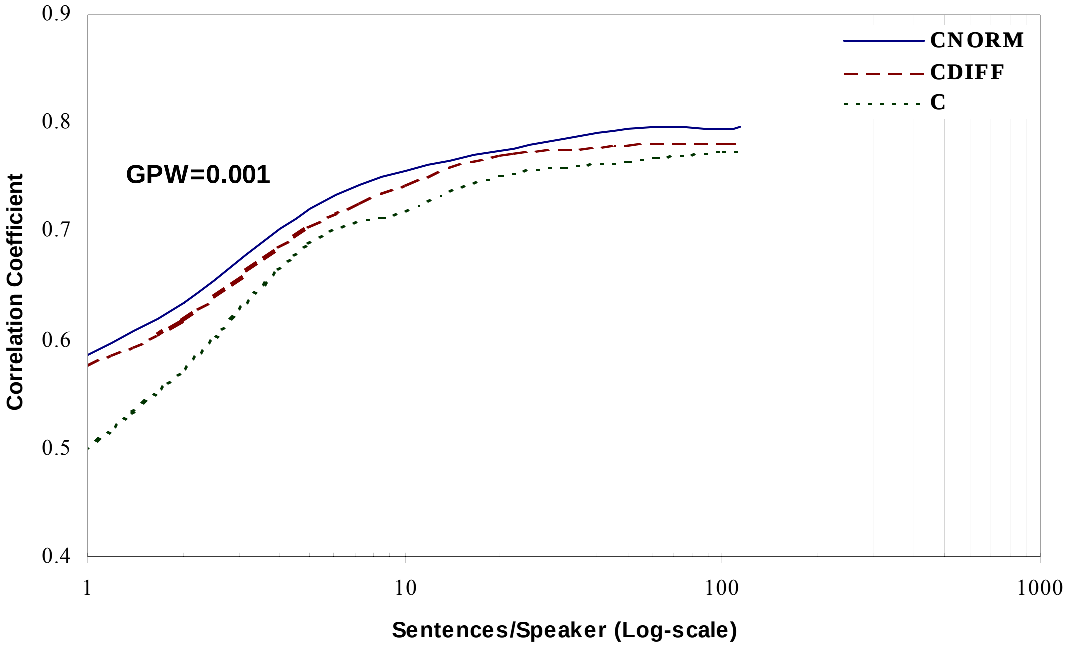

Figure 4: Speaker correlation as function of the amount of sentences for GPW=0.001

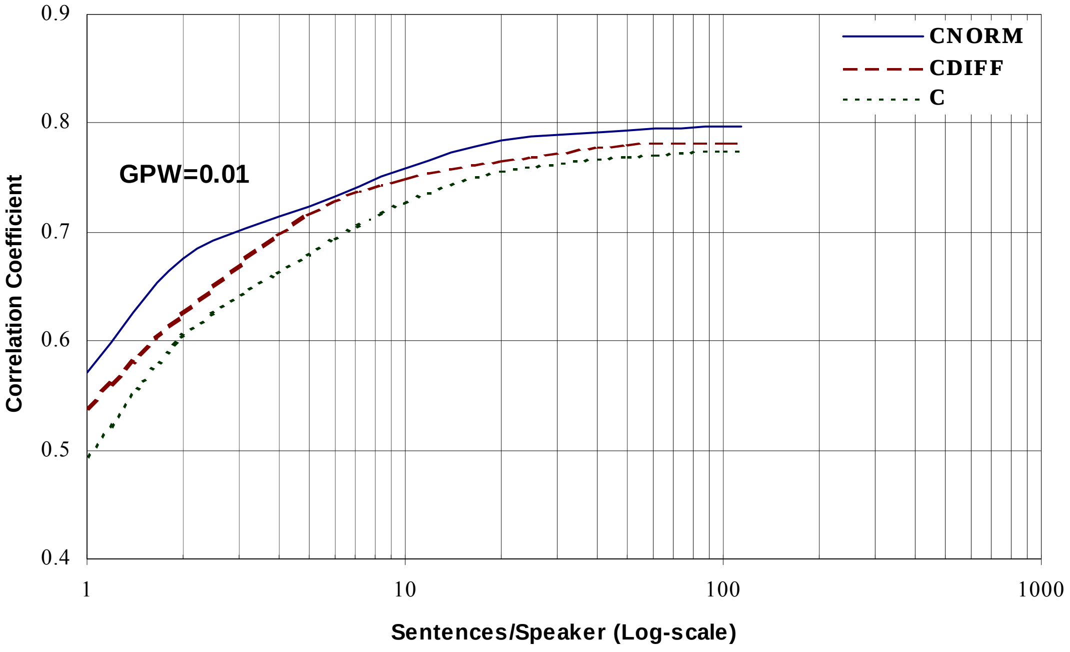

Figure 5: Speaker correlation as function of the amount of sentences for GPW=0.01

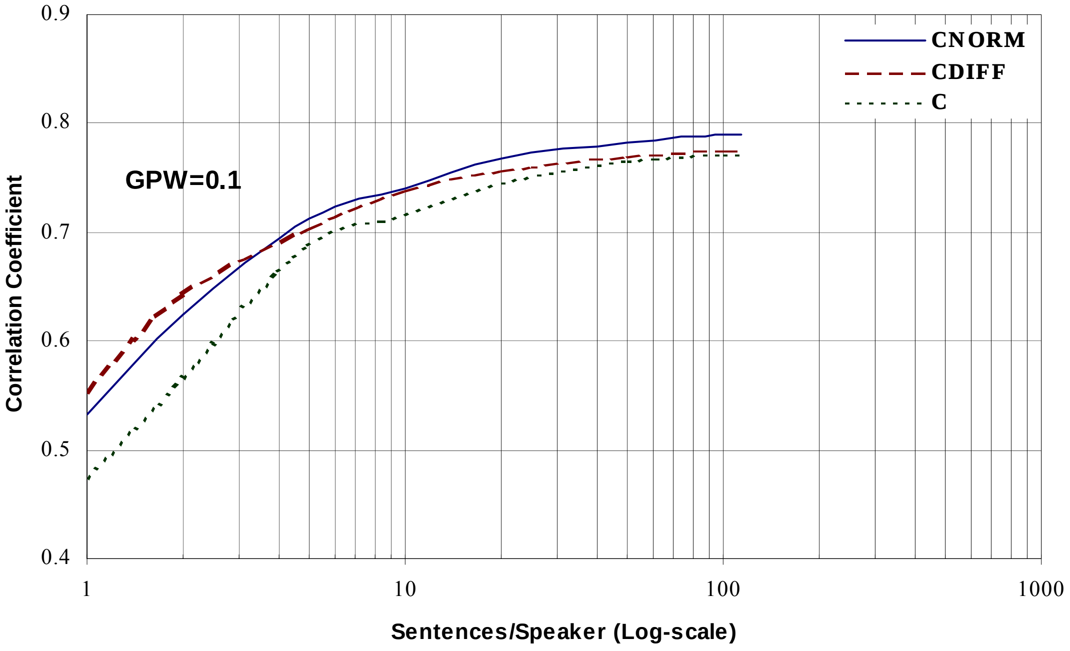

Figure 6: Speaker correlation as function of the amount of sentences for GPW=0.1

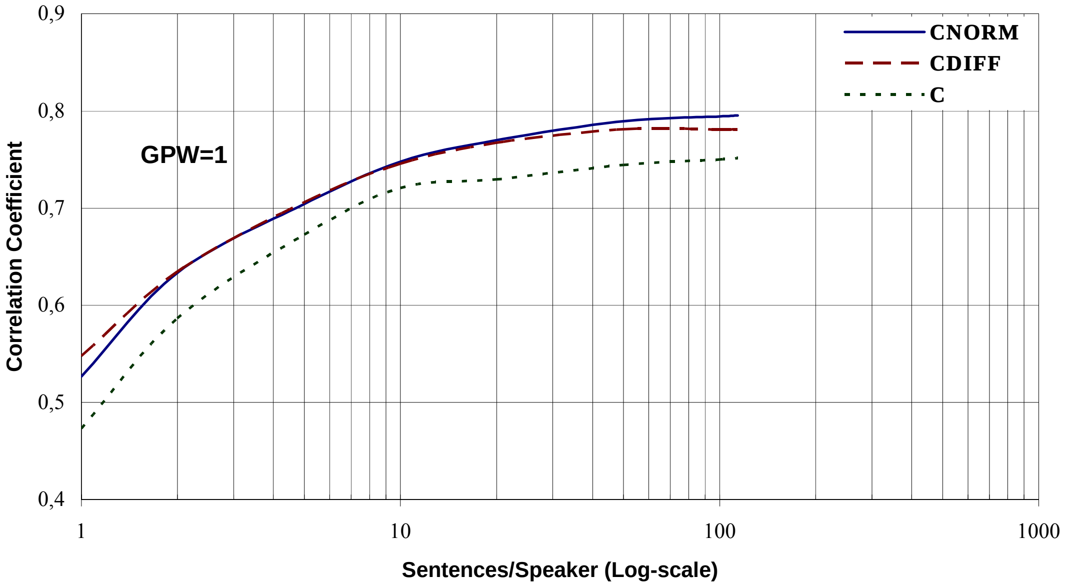

Figure 7: Speaker correlation as function of the amount of sentences for GPW=1