ELSEVIER

Speech Communication 30 (2000) 95–108

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Phone-level pronunciation scoring and assessment for interactive language learning

S.M. Witt *, S.J. Young

Engineering Department, Cambridge University, Trumpington Street, Cambridge CB2 1PZ, UK

Received 20 February 1998; received in revised form 26 November 1998; accepted 2 December 1998

Abstract

This paper investigates a method of automatic pronunciation scoring for use in computer-assisted language learning (CALL) systems. The method utilises a likelihood-based 'Goodness of Pronunciation' (GOP) measure which is extended to include individual thresholds for each phone based on both averaged native confidence scores and on rejection statistics provided by human judges. Further improvements are obtained by incorporating models of the subject's native language and by augmenting the recognition networks to include expected pronunciation errors. The various GOP measures are assessed using a specially recorded database of non-native speakers which has been annotated to mark phone-level pronunciation errors. Since pronunciation assessment is highly subjective, a set of four performance measures has been designed, each of them measuring different aspects of how well computer-derived phone-level scores agree with human scores. These performance measures are used to cross-validate the reference annotations and to assess the basic GOP algorithm and its refinements. The experimental results suggest that a likelihood-based pronunciation scoring metric can achieve usable performance, especially after applying the various enhancements. © 2000 Elsevier Science B.V. All rights reserved.

Zusammenfassung

In diesem Artikel wird eine Methode zur automatischen Bewertung der Aussprache innerhalb eines Systems für computergestütztes Fremdsprachenlernen vorgestellt, welche anhand eines Wahrscheinlichkeitsmaßes, *Goodness of Pronunciation* (*GOP*), einen Aussprachewert für jedes Phoneme in einer Äußerung berechnet. Liegt ein solcher Aussprachewert oberhalb eines Schwellwertes, wurde ein Aussprachefehler detektiert. Die Methode wird im folgendem durch individuelle Schwellwerte für jedes Phoneme, durch die Einbindung von Modellen der Muttersprache des Fremdsprachenschülers und durch Erweiterung der Erkennungsnetzwerke mit zu erwartenden Aussprachefehlern verbessert. Die Evaluation der GOP Methode erfolgt mit Hilfe einer speziell für diese Zwecke aufgenommenen Datenbank englischer Sprache mit ausländischen Akzent, die Phonetikern in Bezug auf Aussprachefehler kommentierten. Da Bewertung von Aussprache höchst subjektiv ist, sind vier Meßmethoden zur Evaluation verschiedener Aspekte der Übereinstimmung verschiedener Bewertungen eines Datensatzes entwickelt worden. Die Anwendung dieser Meßmethoden ermöglicht, die Leistung der GOP Methode mit Phonetikern zu vergleichen. Die experimentiellen Ergebnisse deuten darauf hin, daß eine auf Wahrscheinlichkeitsmetrik zur Aussprachebewertung in der Lage ist, in der Praxis anwendbare Ergebnisse zu liefern; dies gilt insbesondere nach der Anwendung der Verbesserungen. © 2000 Elsevier Science B.V. All rights reserved.

* Corresponding author.
E-mail addresses: smw24@eng.cam.ac.uk (S.M. Witt), sjy@eng.cam.ac.uk (S.J. Young).

0167-6393/00/$ - see front matter © 2000 Elsevier Science B.V. All rights reserved.

PII: S0167-6393(99)00044-8

Keywords: Speech recognition; Computer-assisted language learning; Pronunciation assessment; Pronunciation teaching

1. Introduction

Computer-assisted language learning (CALL) systems can provide many potential benefits to both the language learner and teacher. They allow continuous feedback to the student without requiring the sole attention of the teacher, they facilitate self-study and encourage interactive use of the language in preference to rote-learning. Finally, they can be used to streamline assessment procedures.

To be effective, a CALL system requires the ability to accurately measure pronunciation quality both to enable the immediate correction of errors and to provide longer term feedback on overall language competence. The aim of the work described in this paper is to study acoustic likelihood-based methods for automatic pronunciation assessment within the framework of a hidden Markov model (HMM) speech recognition system.

Existing work on automatic pronunciation scoring has mainly been focussed on the word and phrase level, possibly augmented by measures of intonation, stress and rhythm (Goddijn and de Krom, 1997; Hiller et al., 1993; Hamada et al., 1993; Rogers et al., 1994). These systems typically require several recordings of native utterances to train the models for each word in the teaching material. They are therefore text-dependent with the disadvantage that the teaching material cannot be adjusted without making additional recordings. Systems aimed at teaching selected phonemic errors are described in (Kawai and Hirose, 1997; Kim et al., 1997; Ronen et al., 1997), where either durational information or models trained on non-native speech have been employed. Automatic speech recognition with HMMs has been used to score complete sentences rather than smaller units of speech (Bernstein et al., 1990; Neumeyer et al., 1996). The system described by (Eskenazi, 1996) produces scores for each phone [1] in an utterance, but there is no attempt to relate this to human judgements of pronunciation quality. A dialog system aimed at teaching spoken Japanese is presented in (Ehsani et al., 1997) where speech recognition is used to analyse the student's answer at each stage of the dialog.

The system described here is focussed on measuring pronunciation quality of non-native speech at the phone level. The aims are to locate pronunciation errors, to assess how close the pronunciation is to that of a native speaker and to identify systematic differences when compared to a pronunciation dictionary.

The remainder of this paper is organised as follows. In Section 2, the basic Goodness of Pronunciation (GOP) scoring algorithm is described along with a number of refinements. Section 3 then presents a set of four performance measures which can be used both to validate pronunciation assessments made by human judges and to assess the performance of the GOP algorithms. Section 4 describes the non-native database which was specially recorded for this work. Finally, Sections 5 and 6 present performance assessments of the human judges who annotated the database and the automatic GOP algorithms. The paper concludes with a discussion of the results and some comments on future directions.

2. GOP scoring

2.1. Basic GOP algorithm

The aim of the GOP measure is to provide a score for each phone of an utterance. In computing this score it is assumed that the orthographic transcription is known and that a set of HMMs is available to determine the likelihood $p(O^{(q)}|q)$ of the acoustic segment $O^{(q)}$ corresponding to each phone q. Under these assumptions, the quality of pronunciation for any phone p is defined to be the duration normalised log of the posterior

[1] Throughout this paper a "phone" denotes a sound unit used to model speech with HMMs, which roughly corresponds to a phoneme as defined by linguists.

probability $P(p|O^{(p)})$ that the speaker uttered phone p given the corresponding acoustic segment $O^{(p)}$. That is,

$$\mathrm{GOP}_1(p) \equiv \left|\log\left(P(p|O^{(p)})\right)\right|/\mathrm{NF}(p), \quad (1)$$

$$= \left|\log\left(\frac{p(O^{(p)}|p)P(p)}{\sum_{q\in Q} p(O^{(p)}|q)P(q)}\right)\right| \Big/ \mathrm{NF}(p), \quad (2)$$

where Q is the set of all phone models and $\mathrm{NF}(p)$ the number of frames in the acoustic segment $O^{(p)}$.

Assuming all phones are equally likely ($P(p) = P(q)$) and that the sum in the denominator can be approximated by its maximum, the basic GOP measure becomes

$$\mathrm{GOP}_1(p) = \left|\log\left(\frac{p(O^{(p)}|p)}{\max_{q\in Q} p(O^{(p)}|q)}\right)\right| \Big/ \mathrm{NF}(p). \quad (3)$$

The acoustic segment boundaries and the corresponding likelihoods are determined from Viterbi alignments. Firstly, the numerator of Eq. (3) is computed using a forced alignment in which the sequence of phone models is fixed by the known transcription and secondly, the denominator is determined using an unconstrained phone loop. This is the same arrangement as is commonly used in word spotting (Knill and Young, 1994). One difficulty in Eq. (3) is that if a mispronunciation has occurred, it is not reasonable to constrain the acoustic segment used to compute the maximum likelihood phone q to be identical to the assumed phone $O^{(p)}$. Hence, the denominator score is determined by simply summing the log likelihood per frame over the duration of $O^{(p)}$. In practice, this will often mean that more than one phone in the unconstrained phone sequence has contributed to the computation of $\max_{q\in Q} p(O^{(p)}|q)$.

A block diagram of the resulting scoring mechanism is shown in Fig. 1. The front-end feature extraction converts the speech waveform to a sequence of mel-frequency cepstral coefficients (MFCC) and these are used in two recognition passes: the forced alignment pass and the phone recognition pass where each phone can follow the previous one with equal probability. Based on these results, the individual GOP scores are calculated for each phone as defined in the previous equations. Finally, a threshold is applied to each GOP score to reject badly pronounced phones. The choice of threshold depends on the level of strictness required. The selection of suitable thresholds is further discussed in Section 6.

The quality of the GOP scoring procedure described above depends on the quality of the acoustic models used. Since the aim of the GOP measure is to assess pronunciation quality with respect to native speaker performance, it is

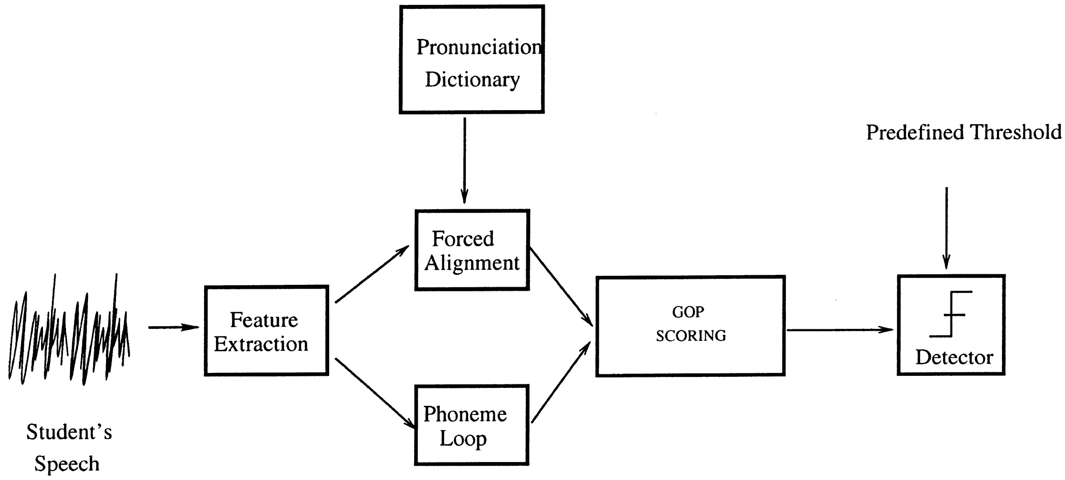

Fig. 1. Block-diagram of the pronunciation scoring system: phones whose scores are above the predefined threshold are assumed to be badly pronounced and are therefore rejected.

reasonable to use native speakers to train the acoustic models. However, non-native speech is characterised by different formant structures compared to those from a native speaker for the same phones (Arslan and Hansen, 1997) and this can lead to phone recognition errors. Hence, some degree of speaker adaptation may be justified. To test this hypothesis, the GOP measure can be computed using models whose Gaussian means have been adapted using Maximum Likelihood Linear Regression (MLLR) (Leggetter and Woodland, 1994). In order to achieve speaker normalisation without adapting to specific phone error patterns, this adaptation is limited to a single global transform of the HMM mixture component means.

2.2. Phone dependent thresholds

So far a single threshold for all phones has been assumed. However, in practice, the acoustic fit of phone-based HMMs differs from phone to phone. For example, fricatives tend to have lower log likelihoods than vowels suggesting that a higher threshold should be used for these.

A simple phone-specific threshold can be computed from the global GOP statistics. For example, the threshold for a phone p can be defined in terms of the mean μ_p and variance σ_p of all the GOP scores for phone p in the training data,

$$T_{p1} = \mu_p + \alpha\sigma_p + \beta, \tag{4}$$

where α and β are empirically determined scaling constants. The assumption here is that averaging the native GOP scores will reduce the affect of errors in the phone recogniser.

A reasonable target for an automatic pronunciation system is to perform as well as a human judge. One way to approximate human performance is to learn from human labelling behaviour. Let $c_n(p)$ be the total number of times that phone p uttered by speaker n was marked as mispronounced by one of the human judges in the training database. Then a second phone dependent threshold can be defined by averaging the normalised rejection counts over all speakers,

$$T_{p_2} = \log \frac{1}{N} \sum_{n=1}^{N} \left(c_n(p) \Big/ \sum_{m=1}^{M} c_n(m) \right), \tag{5}$$

where M is the total number of distinct phones and N the total number of speakers in the training set.

2.3. Explicit error modelling

Pronunciation errors can be grouped into two main error classes. The first class contains individual mispronunciations when a student is not familiar with the pronunciation of a specific word. The second class consists of substitutions of native sounds for sounds of the target language, which do not exist in the native language. This error type will be called systematic mispronunciations. Because the GOP method described so far does not employ models of a student's native phones, incorrect acoustic modelling of the non-native speech will occur especially in the case of systematic mispronunciations. The detection of these errors might be improved if knowledge of the native tongue of the learner can be included in the GOP scoring.

For this purpose a recognition network has been implemented incorporating both correct pronunciation and common pronunciation errors in the form of error sublattices for each phone, using phone model sets of both the target and the native language. Concatenating these sublattices according to target transcriptions yields the desired error network for any utterance. For example, Fig. 2 shows the resulting network for the word "but". The list of possible errors of a Spanish speaker learning English has been taken from (Kenworthy, 1987), some examples of which

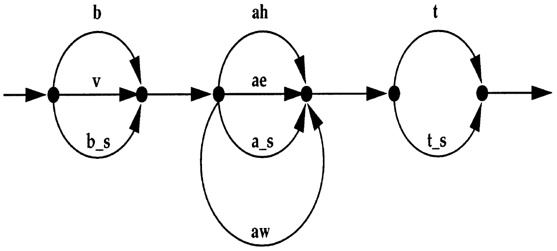

Fig. 2. Example of error network for the word 'but', created through concatenating the sublattice of possible errors for each phone, the topmost phones correspond to the target transcription. (Phoneme names with subscript 's' denote Spanish models.)

Table 1
Expected errors of a Spanish speaker for some British-English phones (Phone names with subscript 's' denote Spanish models)

British	Expected errors
b	b, v, b_s
d	dh, d_s
th	f, s, f_s, s_s
s	hh, del
ch	ch_s
jh	ch,ch_s
k	del, k_s
l	l_s
ah	a_s, aw, ae
uh	uw, u_s
ae	eh, e_s
oh	o_s

are listed in Table 1. The recognition output of such an error network will be a sequence of phones either corresponding to the target transcription $p = p_t$ in the case that the target pronunciation was more likely or otherwise to an error phone $p = p_e$.

A straightforward detector of systematic mispronunciations based on an error network could consist of rejecting all phone segments where an error phone has been recognised. However, such an approach would ignore the information about the likelihood of the occurrence of such an error. Hence, the posterior likelihood of each error phone $P(p_e|O^{(p)})$ is computed by normalising with the recognition results of a phone loop network including acoustic models of both the target language and the speaker's native language applying Eq. (3).

Knowledge of $P(p_e|O^{(p)})$ permits calculating the posterior probability of the target phones p_t in all phone segments containing systematic mispronunciations:

$$\begin{aligned} P(p_t|O^{(p)}) &= 1 - \sum_{q \neq p_t} P(q|O^{(p)}) \\ &\approx 1 - \max_{q \neq p_t} P(q|O^{(p)}) \\ &= 1 - P(p_e|O^{(p)}). \end{aligned} \tag{6}$$

Again the assumption has been made that the above sum can be approximated by its maximum. Thus, scores for systematic mispronunciations $\mathrm{GOP_e}(p)$ are defined as

$$\mathrm{GOP_e}(p) = \begin{cases} |\log(1 - P(p_e|O^{(p)}))| & \text{if } p = p_e, \\ 0.0 & \text{otherwise.} \end{cases} \tag{7}$$

Combining the basic $\mathrm{GOP_1}$ with $\mathrm{GOP_e}$ yields a second GOP metric which includes additional penalties for scores of phone segments where systematic error was recognised.

$$\mathrm{GOP_2}(p) = \mathrm{GOP_1}(p) + K\,\mathrm{GOP_e}(p), \tag{8}$$

where K is a scaling constant.

3. Performance measures

In order to assess the effectiveness of the GOP scoring for detecting pronunciation errors, a set of four new performance measures has been designed. These are based on similarity measurements between reference transcriptions produced by human judges and the output of the GOP metric. Since the production of reference transcriptions must be done by human judges and is highly subjective, the same performance measures are also used to cross-validate the judges. Note that the performance measures are only concerned with the detection of pronunciation errors. They do not take account of the type of error which has occurred.

To cover all aspects of performance, four different dimensions are considered.

- *Strictness* – how strict was the judge in marking pronunciation errors?
- *Agreement* – what is the overall agreement between the reference transcription and the automatically derived transcription? This measure takes account of all phones whether mispronounced or not.
- *Cross-correlation* – what is the overall agreement between the errors marked in the reference and the automatically detected errors? This measure only takes account of phones for which an error has been marked in one or both transcriptions.
- *Overall phone correlation* – how well do the overall rejection statistics for each phone correlate between the reference and the automatic system?

The next section describes the form of the error transcriptions in more detail and then the measures themselves are defined.

3.1. The transcription of pronunciation errors

The non-native database used for assessment consists of target transcriptions based on a pronunciation dictionary and transcriptions which have been annotated by human judges to contain the phone sequence actually spoken. The utterance transcriptions marked with corrections will be referred to as *corrected transcriptions* and the transcriptions derived directly from the pronunciation dictionary will be referred to as the *dictionary-based transcriptions*. Finally, transcriptions in which each phone correction has been replaced by a single rejection symbol are referred to as *rejection-marked transcriptions*.

Two corrected transcriptions of the same utterance are difficult to align with each other due to insertions and deletions of phones. Therefore, all performance measures compare transcriptions on a frame by frame basis. With this approach, measuring the similarity of two differently corrected transcriptions of the same utterance becomes equivalent to comparing the rejection/acceptance marking of corresponding speech frames.

Based on the rejection-marked transcriptions, the frame level markings are calculated as follows:

1. The phone level segmentation for each sentence is calculated by forced alignment of the acoustic waveform with the corrected transcriptions.
2. All frames corresponding to substituted, inserted or deleted phones are marked with “1”, all other ones with “0”. This yields a vector $\boldsymbol{x}$ of length N with $x(i) \in \{0, 1\}$. These vectors will be called transcription vectors.
3. The transitions between “0” and “1” in the transcription vectors are abrupt whereas in practice the precise location of the boundaries between correctly and incorrectly pronounced speech segments are uncertain. Moreover, segmentation based on forced alignments can be erroneous due to the poor acoustic modelling of non-native speech. For these two reasons, the vectors representing corrected transcriptions are smoothed by a Hamming window

$$x'(n) = \sum_{k=-N/2}^{N/2} x(k)w(n-k). \tag{9}$$

Using a frame period of 10 ms, the length of a vowel tends to extend over 6–20 frames, whereas consonants can be much shorter. Also, if rejected frames in one transcription are immediately followed by rejected frames in the other transcription, the rejections can be considered to have been caused by the same pronunciation error. Based on these considerations, a window length of $N = 15$ was selected for all experiments. The effect of the smoothing window is illustrated in Fig. 3.

3.2. Performance measures

This section defines the four performance measures used to compare transcriptions corrected by two judges or one judge and the automatic GOP scoring system.

Firstly, human correction of the pronunciation of non-native speakers depends on personal judgement. There will always exist a large number of phones whose pronunciation is on the borderline between correct and incorrect, a stricter judge might declare more borderline cases as incorrect than another judge who is more benign. In the case of computer-based scoring, the choice of a rejection threshold determines how strict the scoring system will be. This *strictness of labelling*, *S*, can be defined as the overall fraction of phones which are rejected,

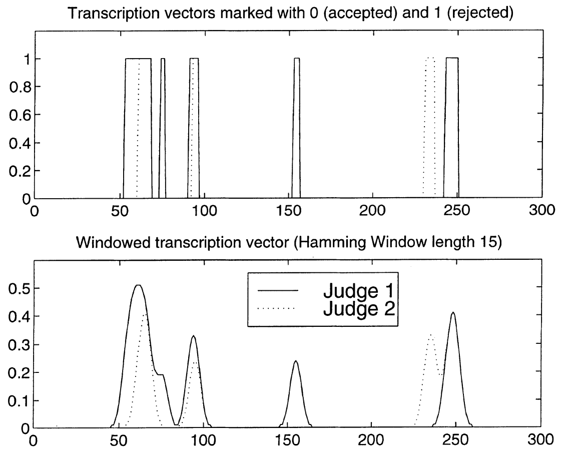

Fig. 3. Smoothing effect of the windowing, overlapping regions denote areas where both judges decided to reject the pronunciation.

$$S = \frac{\text{Count of rejected phonemes}}{\text{Total count of phonemes}}. \tag{10}$$

As an example, the database used for assessment (described below) contains a set of calibration sentences which were labelled by six different judges. Fig. 4 shows the strictness of the judges for these calibration sentences where the mean and standard deviation are $\mu_S = 0.18$ and $\sigma_S = 0.05$, respectively.

A simple way to compare the strictness of two judges $J1$ and $J2$ is to use the difference between strictness levels for the two, i.e.,

$$\delta_S = |S_{J1} - S_{J2}|. \tag{11}$$

The overall *Agreement* (A) between two utterances is defined in terms of the cityblock distance between the corresponding transcription vectors, i.e.,

$$A_{J1,J2} = 1 - \frac{1}{N} \| \boldsymbol{x}_{J1} - \boldsymbol{x}_{J2} \|_{\mathrm{C}}, \tag{12}$$

where $\|\boldsymbol{x}\|_{\mathrm{C}} = \sum_{i=0}^{N-1} |x(i)|$.

Agreement measures overall similarity of two transcriptions by comparing all frames of an utterance. In contrast, the *Cross-Correlation* (CC) measure takes into account only those frames where there exists a rejection in either of them,

$$\mathrm{CC}_{J1,J2} = \frac{\boldsymbol{x}_{J1}^{\mathrm{T}} \boldsymbol{x}_{J2}}{\|\boldsymbol{x}_{J1}\|_{\mathrm{E}} \|\boldsymbol{x}_{J2}\|_{\mathrm{E}}}, \tag{13}$$

where $\|\boldsymbol{x}\|_{\mathrm{E}} = \sqrt{\sum_{i=0}^{N-1} x(i)^2}$ is the standard Euclidean distance. Cross-correlation measures the similarity between all segments which contain rejections in either of the two transcriptions. Because similarity of the rejection patterns with a human judge is the main design objective of the GOP scoring system, this measure has the highest importance.

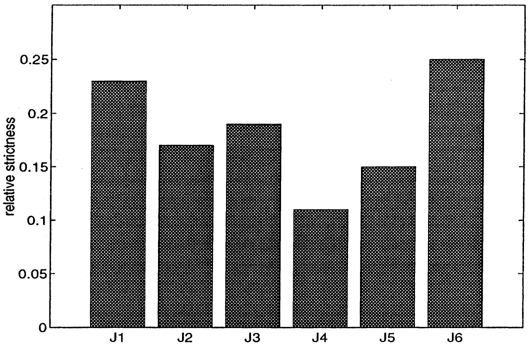

Fig. 4. Relative strictness for all human judges measured on the calibration sentences.

Finally, *Phoneme Correlation* (PC) measures the overall similarity of the phone rejection statistics. Given a vector $\boldsymbol{c}$ whose elements contain the count of rejections for each phone in a complete speaker set, phone correlation is defined as

$$\mathrm{PC}_{J1,J2} = \frac{\sum_{m=0}^{M} (c_{J1}(m) - \mu_{cJ1})(c_{J2}(m) - \mu_{cJ2})}{\sum_{m=0}^{M} \sqrt{(c_{J1}(m) - \mu_{cJ2})^2 (c_{J1}(m) - \mu_{cJ2})^2}}, \tag{14}$$

where μ_c denotes the mean rejection counts.

4. Collection of a non-native database

In order to evaluate the pronunciation scoring methods described above, a database of non-native speech from second-language learners has been recorded and annotated.

The recording guidelines for this database collection were based on the procedures used for the WSJCAM0 corpus (Fransen et al., 1994). Students of English as a second language read prompting texts composed of a limited vocabulary of 1500 words in a quiet room environment. The competence level of the speakers was intermediate. They were all able to understand the prompting texts and instructions and they were able to read the sentences with relatively few hesitations. On the other hand, their competence level was low enough to ensure that they produced a significant number of easily detectable mispronunciations.

Each prompting session consisted of a common set of 40 phonetically balanced sentences read by all subjects and an additional 80 sentences which varied from session to session. Extracts from "Penguin Readers" (Fine, 1995; Chandler, 1991) were used as the source of the prompting texts. These texts have been specifically written for the purpose of teaching English as a foreign language. They employ a limited vocabulary and simplified sentence and grammatical structures.

The subjects for the database collection consisted of 10 students, six of them female and four of them male, speaking as their mother-tongues Latin-American Spanish, Italian, Japanese and Korean. Additionally, 20 sentences from a female Spanish speaker were recorded to serve as calibration sentences. These were annotated by all six judges participating in the labelling task. The resulting six sets of calibration transcriptions can thus be used to compare the human labelling consistency.

The database was annotated by trained phoneticians (the "judges"), all of them native speakers of British English. This annotation was performed at three different levels. Firstly, the original transcriptions were annotated with all substitution, deletion and insertion errors made by the non-native speaker. Since the non-native speech contained a range of sounds which do not exists in British English, the judges had the opportunity to extend the supplied British English phone set with phones taken from the speaker's native language. These labelling instructions yielded transcriptions resembling the non-native speech as closely as possible. Secondly, each word was scored on a scale of 1–4, with 1 representing barely intelligible speech and 4 representing native-like pronunciation. Finally, each sentence was scored on the same scale. Of these three levels of annotation, only the first phone error correction level is used for the experiments reported here.

5. The labelling consistency of the human judges

In order to properly interpret the results of assessing a computer-based pronunciation system using manually derived transcriptions as the reference, it is necessary to measure the inter-judge labelling consistency and to obtain an understanding of how the judges label the data. Their labelling is characterised by the phones they consider important for good pronunciation and thus tend to correct, by the consistency of the rejection patterns across different judges and finally by their strictness. In this section, the four performance measures described above are used in conjunction with the 20 calibration sentences to determine these characteristics.

Fig. 5 shows averaged results of all the measures for each judge. These results have been calculated by averaging A, CC, PC and δ_S between the respective judge and all other ones. All results vary within an acceptable range, that is $0.85 < A < 0.95$, $0.40 < \mathrm{CC} < 0.65$, $0.70 < \mathrm{PC} < 0.85$ and $0.03 < \delta_S < 0.14$. Therefore, the labelling by different human judges can be considered as being reasonably consistent although Judge 5 is a slight

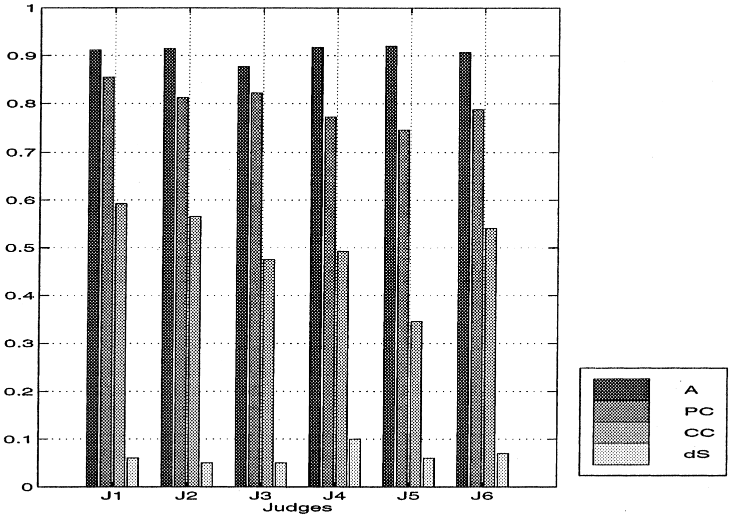

Fig. 5. A, CC, PC and δ_S for each judge based on averaging the measures between the respective judge and all the other judges.

outlier in that he has a lower average cross-correlation with the other judges. The total mean values over all pairs of judges of all four measures are shown in Table 2. These mean values will be used as benchmark values against which to measure the performance of the automatic scoring presented in later sections.

Table 3 shows the similarity between the human judges and the baseline GOP scoring method for each non-native speaker in that judge's group. It can be seen that the intra-judge results are quite consistent. However, Judge 4 had a high acceptance of non-native pronunciation, and thus corrected a significantly smaller portion of the data. For such a degree of strictness, the automatic scoring performs considerably worse.

Fig. 6 shows the CC and PC measures for each speaker grouped according to their native languages. Also shown on this figure are the genders of each speaker. From this figure and the data shown in Fig. 5, it appears that the labelling of the human judges does not depend significantly on the mother tongue or the gender of the subjects, but depends mostly on the variability of human judges.

Table 2
Averaged A, CC, PC and δ_S results based on correlating all possible pairs of judges (these values are the baseline against which to measure automatic scoring performance)

A	CC	PC	δ_S
0.91	0.47	0.78	0.06

Table 3
Similarity results between judges and the baseline GOP scoring grouped according to the judge who labelled the respective speaker sets (the speaker name Cal. denotes the calibration sentences)

Judge	Speaker	Strictness	CC	PC
J1	Cal.	0.25	0.51	0.77
	ss	0.25	0.56	0.73
	ts	0.21	0.49	0.84
J2	Cal.	0.19	0.53	0.81
	yp	0.16	0.49	0.62
J3	Cal.	0.21	0.50	0.68
	mk	0.13	0.38	0.57
J4	Cal.	0.13	0.37	0.62
	sk	0.07	0.12	0.61
	as	0.11	0.37	0.50
J5	Cal.	0.16	0.22	0.71
	ay	0.19	0.50	0.61
	fl	0.16	0.43	0.56
	pc	0.19	0.50	0.62
	ky	0.23	0.48	0.34

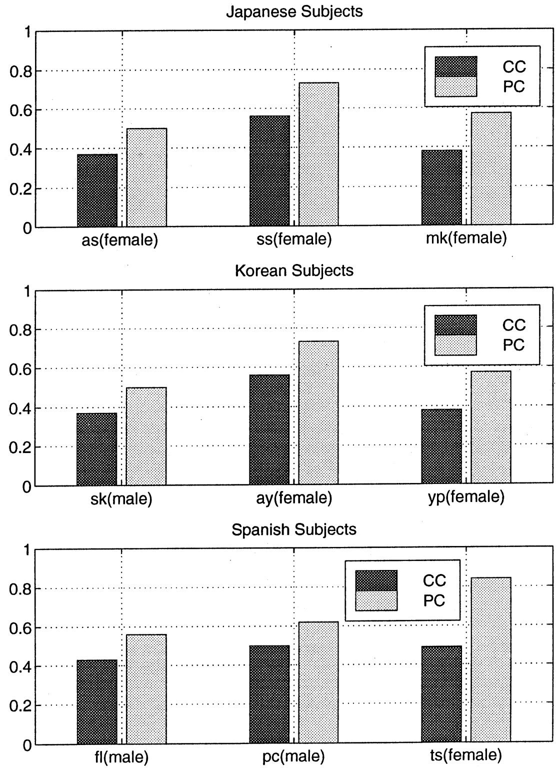

Fig. 6. CC and PC results grouped according to each student's mother-tongue.

Finally, the rejection patterns for three of the judges are shown in Fig. 7, which depicts the rejection counts for all phones for the judges. The strong correlation between the rejection pattern of these three judges is clearly evident.

The above analysis of human judgement characteristics shows that although there is significant variability in the labelling of each judge, there is nevertheless sufficient common ground to form a basis for assessing the performance of the various automatically derived pronunciation metrics.

6. Experimental results

This section presents performance results for both the basic GOP scoring method and the

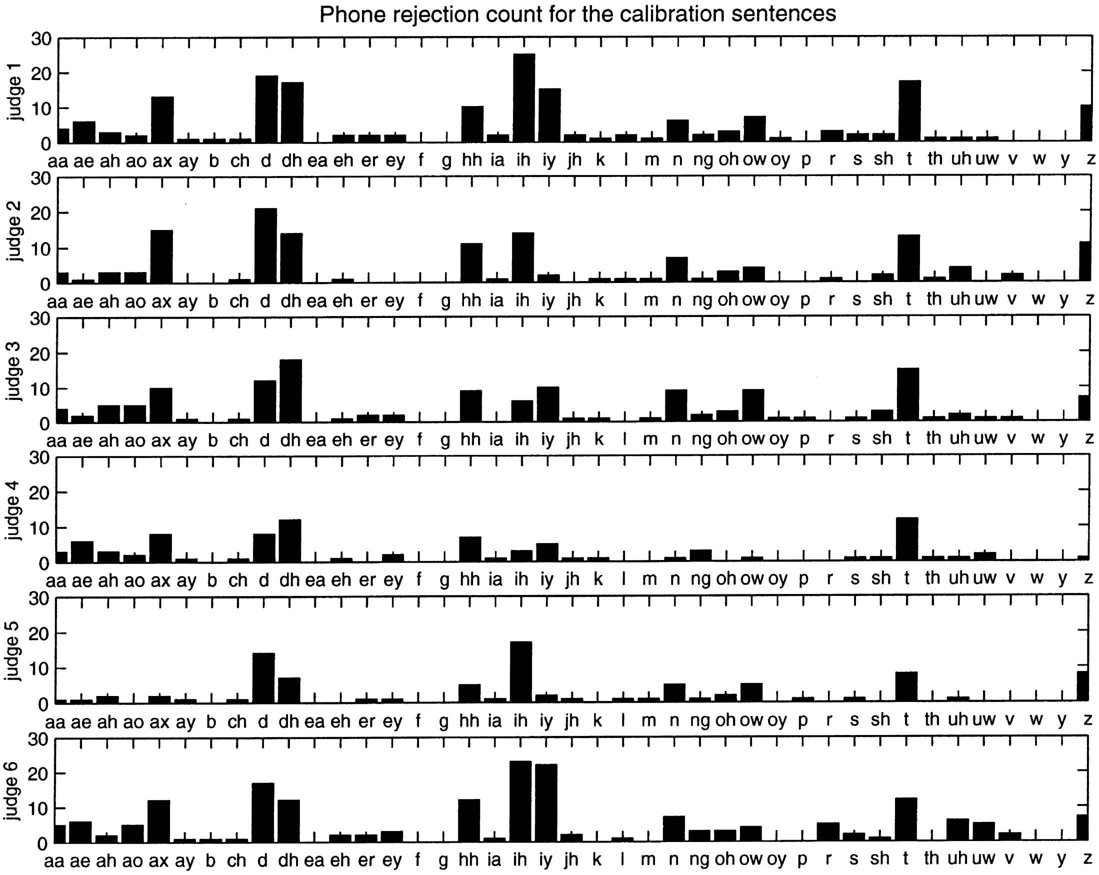

Fig. 7. Rejection counts of all phones for all judges based on the calibration sentences to show the correlation between the rejection pattern of different judges.

various refinements described in Section 2. All speech recognition is based on multiple mixture monophone models trained on the British English corpus WSJCAM0 (Fransen et al., 1994). The HMMs were built using the HTK Toolkit (Young et al., 1996).

For the case of automatic GOP scoring, agreement A, cross-correlation CC and phone correlation PC vary according to the level of strictness applied, which again depends on the threshold levels set. In Fig. 8, the GOP scores for an example sentence are shown. Varying the threshold determines the number of rejections. For the marked threshold in the example figure, both human and machine judgements agree on which phones to accept and to reject with two exceptions. The first phone of the sentence is not rejected by the human judge but it is rejected by the GOP metric, this is probably due to bad acoustic modelling at the sentence beginning. Further, the 'ae' in 'carry' has been rejected by the human judge method but not by the GOP scoring.

In the work reported here, the range of rejection thresholds studied was restricted to lie within one standard deviation of the judges strictness i.e., $|\delta_S| \leqslant \sigma_S$ where in this case $\sigma_S = 0.05$. Within this range, the variation of A, CC and PC for one speaker as a function of the threshold level is shown in Fig. 9. In this figure, the vertical lines denote the acceptable range of threshold settings and, as can be seen, the performance values do not vary greatly within this range.

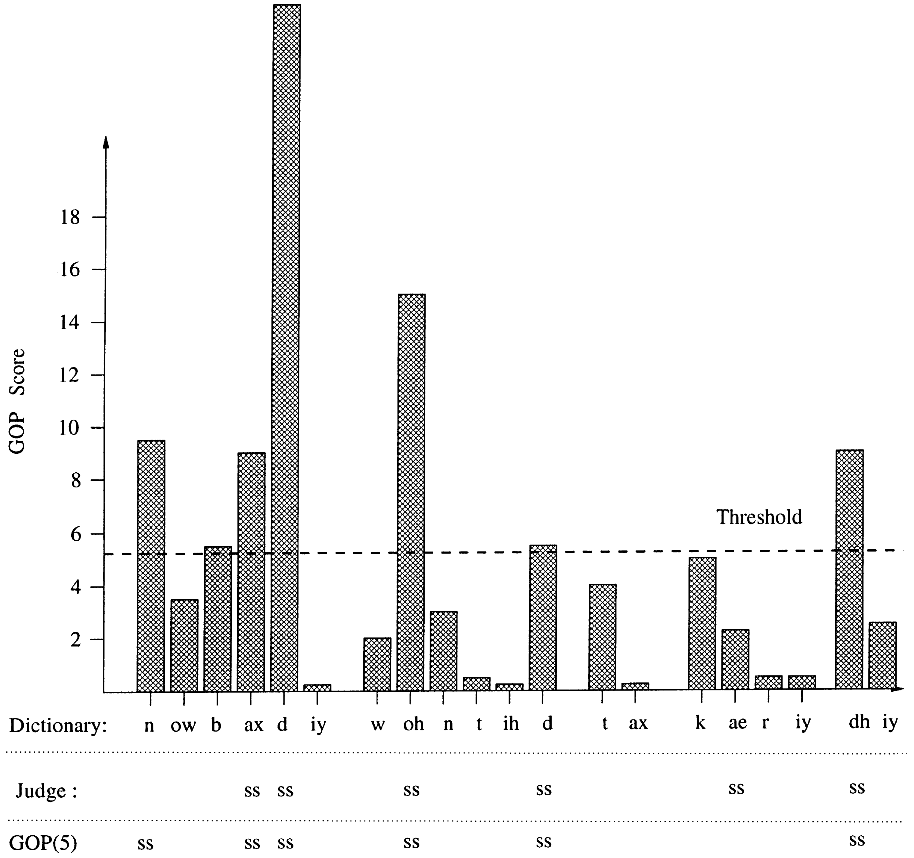

Fig. 8. GOP scoring results for an example sentence, 'ss' denotes the location of a rejection, the automatically rejected phones correspond to GOP scores above the threshold.

Table 4 shows optimal values of A, CC and PC achievable for each speaker within the allowed threshold range. As can be seen, the optimal threshold is speaker dependent. However, apart from speakers 'sk', 'as', a threshold of 4.5 would be close to optimal for all speakers. Since 'sk' and 'as' were the two speakers whose transcriptions were annotated by the very strict judge (Judge 4), these two speakers have not been included in the results presented in the following.

The performance results for the automatic GOP scoring metrics as discussed in Section 2 are summarised in Fig. 10. The first bar on the left marked "Baseline" shows the performance of the basic GOP 1 metric with a fixed overall threshold as discussed above. The final bar on the left shows the human–human performance on the calibration sentences for comparison. As can be seen, the scores for A and CC are similar whereas for PC, the automatic scoring is worse by about 20%. The second bar marked "MLLR" shows the effect of applying speaker adaptation. For Group 1 an improvement of 5% has been obtained for PC at the cost of a small decrease in CC. The third and fourth bars show the effects of using individual thresholds for each phone based on averaging native scores T_{p_1} and on averaging the judges scores T_{p_2}. As can be seen, thresholds derived from the statistics of the judge's scoring appear to provide the best performance. This is probably because these are directly related to the desired rejection statistics.

Finally, Table 5 summarises the effects of incorporating error modelling into the GOP

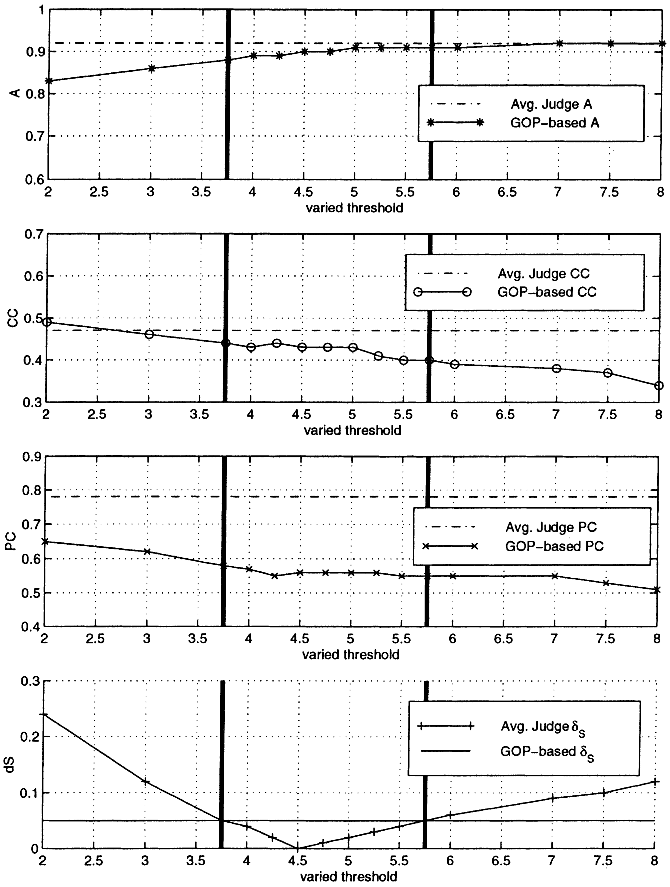

Fig. 9. Dependency of A, CC, PC and δ_S on threshold variation, based on data for 'fl', a male Spanish speaker. The range inside the bold lines is the range of valid δ_S.

algorithm. The British English HMMs were augmented by a set of similar Spanish models trained on a database of Latin-American Spanish. The data from the three Spanish speakers in the database was analysed using the extended GOP_2 metric with the scale factor K adjusted to give optimal performance. Instead of averaging over all eight speakers, results shown in this table are only averaged over the three Spanish speakers in the database, the averaged baseline performance of which is shown in the first line of Table 5. Comparision of these results with those for the second GOP metric demonstrate that a slight improvement can be obtained by including the extra information relating to systematic mispronunciations. Finally, the results of combining all proposed refinements of the baseline algorithm, i.e., adaptation, judge-based individual thresholds and error modelling, are as high as the human benchmark values.

Table 6 compares human rejections with all automatically detected systematic mispronunciations, i.e., all rejections of phone segments where a native phone had been more likely than the target phone. The relatively high values for CC and PC indicate that a large proportion of pronunciation errors are due to systematic mispronunciations and that a significant proportion of these can also be detected by the use of error networks. Additionally, this metric provides information about which type of mispronunciations occurred and whether the pronunciation of a given phoneme sounds more Spanish than English. This

Table 4
Thresholds yielding optimal performance for all non-native speakers of the database

ID	Thres	A	CC	PC	δ_S
fl	5	0.91	0.43	0.56	0.02
pc	4.5	0.87	0.50	0.62	0.04
yp	4.0	0.90	0.49	0.62	0.02
ts	4	0.87	0.49	0.84	0.03
ky	5	084	0.48	0.34	0.04
sk	7	0.90	0.12	0.61	0.06
ss	4.5	0.85	0.56	0.73	0.05
as	7	0.90	0.37	0.50	0.07
mk	4.5	0.90	0.38	0.57	0.07
ay	4.5	0.90	0.50	0.61	0.05
j1	4.5	0.86	0.51	0.77	0.01
j2	5	0.87	0.53	0.81	0.04
j3	4.5	0.86	0.50	0.68	0.05
j4	7.5	0.91	0.43	0.62	0.00
j5	5.5	0.88	0.22	0.71	0.05
j6	4	0.85	0.52	0.65	0.03
GOP mean		0.88	0.47	0.60	0.05
Human mean		0.91	0.47	0.78	0.05

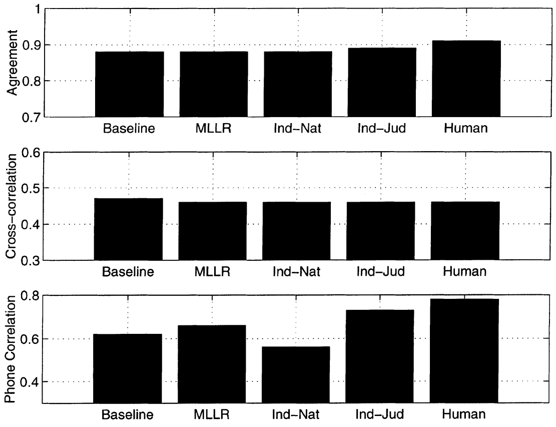

Fig. 10. Comparison of the A, CC and PC performance measures using: (a) the basic GOP scoring (Baseline); (b) basic GOP with adaptation (MLLR); (c) individual thresholds based on native scores (Ind-Nat); (d) individual thresholds based on human judges (Ind-Jud); (e) human–human average performance (Human).

information might be used in future work to provide additional feedback about error types in addition to detecting error locations within an utterance or a word.

Table 5
Performance for experiments with and without error modelling (all experiments include adaptation)

Experimental setup	A	CC	PC
Baseline	0.89	0.46	0.71
Ind-Judge	0.89	0.48	0.76
Error modelling	0.88	0.48	0.72
Ind-Judge + Error modelling	0.90	0.49	0.78
Human mean	0.91	0.47	0.78

Table 6
Performance results when using an error network to detect systematic mispronunciations

Speaker ID	A	CC	PC	S_{ID}	δ_S
fl	0.90	0.34	0.42	0.18	0.01
pc	0.88	0.39	0.57	0.23	0.04
ts	0.89	0.27	0.40	0.19	0.02

7. Conclusions

This paper has presented a likelihood-based method for GOP scoring and proposed a number of metrics which can be used to assess performance in comparison to human judges. Using a specially recorded database of non-native speech, the basic GOP method has been investigated and the effectiveness of the performance measures studied.

The combination of the baseline method with several refinements yielded improvements in the automatic scoring performance, which then became comparable to the human–human benchmark values. Applying speaker adaption and individual thresholds trained on human judgements has improved the phone correlation from PC = 0.62 to 0.72, this being only about 7.7% worse than the averaged human performance of PC = 0.78. For the Spanish speaker sets in the database, application of the error modelling technique yielded performance as high as the benchmark values.

In conclusion, this work indicates that a computer based pronunciation scoring system is likely to be capable of providing similar feedback to a student as a human judge with regard to which phonetic segments in an utterance can be accepted as correct or not. Future work will concentrate on expanding the algorithm to inform the student about which mistake he or she has made.

References

Arslan, L., Hansen, J., 1997. Frequency characteristics of foreign accented speech. In: ICASSP'97. Munich, Germany.

Bernstein, J., Cohen, M., Murveit, H., Ritschev, D., Weintraub, M., 1990. Automatic evaluation and training in English pronunication. In: ICSLP'90. Kobe, Japan, pp. 1185–1188.

Chandler, R., 1991. Farewell, My Lovely. Penguin Readers Series, Penguin, Harmondsworth.

Ehsani, F., Bernstein, J., Najmi, A., Todic, O., 1997. Subarashi: Japanese interactive spoken language education. In: Proceedings EUROSPEECH'97. Rhodes, Greece.

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

Fine, A., 1995. Madame Doubtfire. Penguin Readers Series, Penguin, Harmondsworth.

Fransen, J., Pye, D., Robinson, A., Woodland, P., Young, S., 1994. WSJCAM0 corpus and recording description. Technical Report CUED/F-INFENG/TR 192, Cambridge University Engineering Department, Cambridge, UK.

Goddijn, A., de Krom, G., 1997. Evaluation of second language learners' pronunciation using hidden Markov models. In: Proceedings EUROSPEECH'97. Rhodes, Greece.

Hamada, H., Miki, S., Nakatsu, R., 1993. Automatic evaluation of English pronunciation based on speech recognition techniques. IEICE Trans. Inform. Syst. E76-D (3), 352–359.

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

Kawai, G., Hirose, K., 1997. A call system using speech recognition to train the pronunciation of japanese long vowels, the mora nasal and mora obstruent. In: Proceedings EUROSPEECH'97. Rhodes, Greece.

Kenworthy, J., 1987. Teaching English Pronunciation. Longman, New York.

Kim, Y., Franco, H., Neumeyer, L., 1997. Automatic pronunciation scoring of specific phone segments for language instruction. In: Proceedings EUROSPEECH'97. Rhodes, Greece.

Knill, K., Young, S., 1994. Speaker dependent keyword spotting for accessing stored speech. Technical Report CUED/F-INFENG/TR 193. Cambridge University Engineering Department, Cambridge, UK.

Leggetter, C., Woodland, P., 1994. Speaker adaptation of HMMs using linear regression. Technical Report CUED/F-INFENG/TR. 181. Cambridge University Engineering Department, Cambridge, UK.

Neumeyer, L., Franco, H., Weintraub, M., Price, P., 1996. Pronunciation scoring of foreign language student speech. In: ICSLP'96, Philadelphia, PA, USA.

Rogers, C., Dalby, J., DeVane, G., 1994. Intelligibility training for foreign-accented speech: A preliminary study. J. Acoust. Soc. Amer. 96 (4), pt. 2.

Ronen, O., Neumeyer, L., Franco, H., 1997. Automatic detection of mispronunciation for language instruction. In: Proceedings EUROSPEECH'97. Rhodes, Greece.

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