PERFORMANCE MEASURES FOR PHONE-LEVEL PRONUNCIATION TEACHING IN CALL

S.M. Witt S.J. Young
Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ
United Kingdom
Email: {smw24,sjy}@eng.cam.ac.uk

ABSTRACT

This work presents a general development framework for automatic pronunciation assessment within computer-assisted language learning (CALL) together with several refinements of a previously described pronunciation scoring method. This method utilises a likelihood-based 'Goodness of Pronunciation' (GOP) measure which in this work has been extended to include individual thresholds for each phone based on both averaged native confidence scores and on rejection statistics provided by human judges. These statistics where provided through a specifically recorded and annotated database of non-native speech. 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.

1. INTRODUCTION

Computer-assisted language learning (CALL) systems can provide many potential benefits to both the language learner and teacher. Especially, incorporating interactivity in a CALL system enables continuous feedback to the student in a self-study environment. Interactivity with regard to pronunciation teaching in a CALL system requires the ability to accurately measure a student's pronunciation quality in order to enable immediate detection and correction of errors. Previous pronunciation systems aimed at teaching selected phonemic errors are described in [3], [4] and [6], where either durational information or models trained on non-native speech have been employed. The system described by [1] produces scores for each phone [1] in an utterance but there is no attempt to relate this to human judgements of pronunciation quality.

The aim of the work described here is to develop an evaluation framework for any phone-level pronunciation scoring method. Therefore, Section 2 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 an automatic scoring system. Section 3 then presents performance assessments of the human judges who annotated the non-native database. Knowledge derived from sections 2 and 3 has been incorporated in refinements of an acoustic likelihood-based method for automatic assessment of pronunciation quality of non-native speech at the phone level with the aim of locating pronunciation errors described in earlier work [7]. The paper concludes with a discussion of the results and some comments on future directions.

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

2. PERFORMANCE MEASURES

2.1. 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*. Transcriptions in which each phone correction has been replaced by a single rejection symbol are referred to as *rejection-marked transcriptions*. Because phones insertions and deletions make the alignment of two rejection-marked transcriptions of the same utterance difficult, all performance measures compare transcriptions on a frame by frame basis. Thus, the similarity of two differently corrected transcriptions of the same utterance becomes equivalent to comparing the rejection/acceptance marking of corresponding speech frames.

These 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 *transcription vector* $\mathbf{x}$ of length N with $x(i) \in \{0, 1\}$.
3. The abrupt transitions between "0" and "1" in these vectors do not reflect the uncertainty in the precise location of the boundaries between correctly and incorrectly pronounced speech segments. 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 with a window length $N = 15$.

2.2. Performance measures

In order to assess the effectiveness of the GOP scoring for detecting pronunciation errors, a set of four performance measures has been designed. These measures are based on frame-level 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.

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

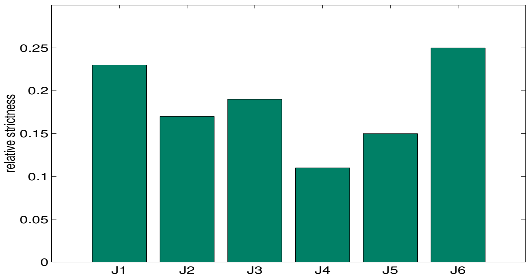

Figure 1. Relative strictness for all human judges measured on the calibration sentences

- *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 agree between the reference and the automatic system?

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

$$S = \frac{Count\ of\ Rejected\ Phones}{Total\ Count\ of\ Phones} \tag{1}$$

As an example, the non-native database (described in [7]) contains a set of calibration sentences labelled by six different judges, for which the strictness is shown in Figure 1, 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.

$$\delta_S = \mid S_{J1} - S_{J2} \mid \tag{2}$$

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

$$A_{J1J2} = 1 - \frac{1}{N} \mid\mid \mathbf{x}_{J1} - \mathbf{x}_{J2} \mid\mid_C \tag{3}$$

where $\mid\mid \mathbf{x} \mid\mid_C = \sum_{i=0}^{N-1} \mid x(i) \mid$.

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

$$CC_{J1,J2} = \frac{\mathbf{x}_{J1}^T \mathbf{x}_{J2}}{\mid\mid \mathbf{x}_{J1} \mid\mid_E \mid\mid \mathbf{x}_{J2} \mid\mid_E} \tag{4}$$

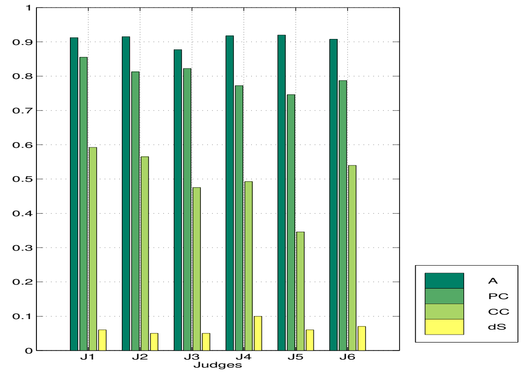

Figure 2. A, CC, PC and δ_S for each judge based on averaging the each measure between the respective judge and all the other judges.

where $\mid\mid \mathbf{x} \mid\mid_E = \sqrt{\sum_{i=0}^{N-1} x(i)^2}$ is the standard Euclidean distance. Cross-correlation measures 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.

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

$$PC_{J1,J2} = \frac{\sum_{m=0}^{M} (c_{J1}(m) - \mu_{c_{J1}})(c_{J2}(m) - \mu_{c_{J2}})}{\sum_{m=0}^{M} \sqrt{(c_{J1}(m) - \mu_{c_{J1}})^2 (c_{J1}(m) - \mu_{c_{J1}})^2}}$$

where μ_c denotes the mean rejection counts.

3. THE LABELLING CONSISTENCY OF THE HUMAN JUDGES

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, see also [7]. In order to measure labelling consistency a set of 20 calibration sentences has been annotated by six human judges. In addition to these, there are 10 sets of non-native speakers each of them containing 120 sentences of clean read speech.

To interpret results of computer-based pronunciation assessment 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. Human 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. The four new performance measures have been used in conjunction with the 20 calibration sentences to determine these characteristics.

Figure 2 shows averaged results of all the measures for each judge. All results vary within an acceptable range, therefore, the labelling by different human judges can be considered as being reasonably consistent although Judge 5 is a slight 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 1. These mean values will be used as benchmark values against which to measure the performance of the automatic scoring presented in later sections.

A	CC	PC	δ_S
0.91	0.47	0.78	0.06

Table 1. 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

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

Table 2. Similarity results between judges and the baseline GOP scoring grouped according to the judge who labelled the respective speaker sets (Spkr). The speaker name (Cal.) denotes the calibration sentences. Additionally, the gender (G) and mother tongue (MT) of each subject are shown.

Table 2 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.

From the data shown in Table 2, 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. This 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.

4. REFINEMENTS OF THE GOODNESS OF PRONUNCIATION (GOP) SCORING

The aim of the GOP measure is to provide a score for each phone p of an utterance. In computing this score it is assumed that the orthographic transcription is known and that a set of hidden Markov models Q is available. Based on the derivation in [7] the basic GOP measure is defined as:

$$GOP_1(p) = \left| \log \left(\frac{p(O^{(p)}|p)}{\max_{q\in Q} p(O^{(p)}|q)} \right) \right| / NF(p) \qquad (5)$$

with NF denoting the number of frames of the utterance segment aligned to the phone p.

The quality of the *GOP* scoring procedure depends on the quality of the acoustic models used. Because the target is to teach native-like speech, models trained on native speakers were employed. However, the different formant structures of non-native speech will often cause phone recognition errors. In order to achieve speaker normalisation without adapting to specific phone error patterns, speaker adaptation based on a single global transform of the HMM mixture component mean using Maximum Likelihood Linear Regression (MLLR) [5] has been implemented.

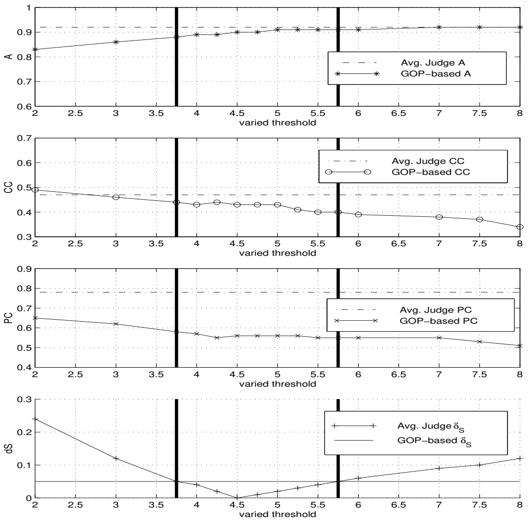

Figure 3. Dependency of A, CC, PC and δ_S on threshold variation, based on data for 'fl', a male Spanish speaker. The range inside the bold lines is the range of valid δ_S

4.1. Phone Dependent Thresholds

The basic algorithm assumes a single threshold for all phones. 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

$$T_{p_1} = \mu_p + \alpha\sigma_p + \beta \qquad (6)$$

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

$$T_{p_2} = log \frac{1}{N} \sum_{n=1}^{N} \left(c_n(p) / \sum_{m=1}^{M} c_n(m) \right) \qquad (7)$$

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

5. EXPERIMENTAL RESULTS

This section presents performance results for both the basic GOP scoring method and the refinements described in section 4. All speech recognition is based on multiple mixture monophone models trained on the British English corpus WSJCAM0 [2] using the HTK Toolkit [8].

For the case of automatic *GOP* scoring, A, CC and PC vary according to the level of strictness applied, which

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	0.84	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
GOP Mean		0.88	0.47	0.60	0.05
Human Mean		0.91	0.47	0.78	0.05

Table 3. Thresholds yielding optimal performance for all non-native speakers of the database

again depends on the threshold levels set. The range of rejection thresholds was restricted to lie within one standard deviation of the judges strictness i.e. $|\delta_S| \leq \sigma_S$ 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 Figure 3. 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.

Table 3 shows optimal values of A, CC and PC achievable for each speaker within the allowed threshold range. These optimal thresholds are 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, see Table 2), these two speakers have not been included in the following averaged results, which are summarised in Figure 4. The first bar on the left marked "Baseline" shows the performance of the basic GOP_1 metric with a fixed overall threshold. 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. 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}. Thresholds derived from the statistics of judge's scoring appear to provide the best performance. This is probably because these are directly related to desired rejection statistics.

6. CONCLUSIONS

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

The refinements of the baseline methods yielded improvements in the automatic scoring performance, which thus becomes close 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 $PC = 0.72$, this being only about 7.7% worse than the averaged human performance of $PC = 0.78$. In conclusion, this work indicates that a computer based pronunciation scoring system is likely to be capable of providing similar feedback to a student than 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.

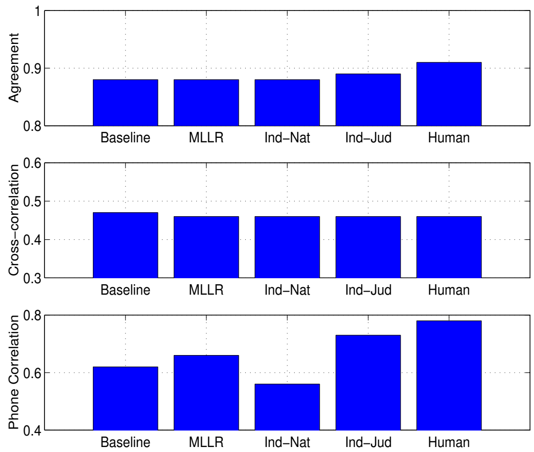

Figure 4. Comparison of the A,CC and PC performance measures using (a) the basic GOP scoring (Baseline), (b) basic GOP with adaptation (MLLR), (c) individual thresholds based on native scores (Ind-Nat), (d) individual thresholds based on human judges (Ind-Jud), and (e) Human-human average performance (Human)

REFERENCES

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

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

[3] G. Kawai and K. Hirose. 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, 1997.

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

[5] C.J. Leggetter and P.C. Woodland. Speaker adaptation of HMMs using linear regression. Technical Report CUED/F-INFENG/TR. 181, Cambridge University Engineering Department, Cambridge, U.K., June 1994.

[6] O. Ronen, L. Neumeyer, and H. Franco. Automatic detection of mispronunciation for language instruction. In *Proceedings EUROSPEECH '97*, Rhodes, Greece, 1997.

[7] S.M. Witt and S.J. Young. Language Learning based on Non-native Speech Recognition. In *Proceedings EUROSPEECH '97*, pages 633–636, Rhodes, Greece, 1997.

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