A Phone Verification Approach to Pronunciation Quality Assessment for Spoken Language Learning

Khe Chai Sim
School of Computing, National University of Singapore, Singapore
E-mail: simkc@comp.nus.edu.sg Tel: +65 6516 4813

***Abstract*—Computer-assisted language learning (CALL) is a form of computer-based assisted learning used in teaching to facilitate the language learning process. One major aspect of CALL for spoken language learning is the automatic assessment of pronunciation quality. It greatly relies on speech recognition technology to provide gradings for the pronunciation quality of the given input speech. This paper introduces a phone verification approach which allows the detection of mispronunciations at phone level. The detection thresholds can be determined based on the equal error rate (EER) metric using a database containing only native speech. In addition this approach also allows aggregation of assessment scores at sentence and speaker levels by computing the average phone rejection rates. This paper compares three different methods of generating goodness of pronunciation confidence scores. In addition, this paper also examines both unsupervised versus supervised adaptation techniques to improve the verification performance. Experimental results are reported based on the EER metric.**

I. Introduction

Computer-assisted language learning (CALL) is a form of computer-based assisted learning used in teaching to facilitate the language learning process. The primary focus of a CALL system is to provide audio and visual interactivity between the learner and computer to enchance the learning experience. The learning materials presented by a CALL system are typically in the form of texts, images, audio and videos. During the learning process, the learner may interact with the computer via keyboard and pointer inputs. In order to assess the spoken skills of the learner, a CALL system needs to be able to perform automatic pronunciation quality assessment through speech input.

The earlier work done on automatic pronunciation assessment is based on generating machine scores at the sentence or speaker level [5]. The performance of these systems were evaluated based on the correlation with the scores annonated by human experts. More recently, Witt and Young [7] introduced a phone level Goodness of Pronunciation (GoP) to detect mispronunciations. According to the paper, phone level posterior (confidence) scores are generated and compared against a threshold to detect mispronunciations. In that paper, test sets were '*artificially*' created by replacing some of the phones in the dictionary with similar sounding phones. Therefore, the locations of mispronunciation can be identified and the detection performance can be evaluated.

This paper proposes a phone verification approach to automatic pronunciation assessment. This system prompts the user to speak a sentence. The system then generates phone-level confidence scores which are used to verify the identity of the phone. Instead of arbitrarily replacing some of the phones to '*simulate*' pronunciation errors, a *complete* verification analysis can be performed by verifying each phone against all the phone models in the system. Typically, the performance of a verification system is evaluated based on the Equal Error Rate (EER) metric, the average false acceptance and false rejection error rates at the operating point where these errors are equal. The EER metric not only measures the performance of the verification system to determine its effectiveness in automatic pronunciation assessment, it also allows the optimum decision threshold to be computed. Given the decision threshold, the verification system provides a phone-level binary feedback to the user indicating whether the phones have been pronounced correctly (accepted) or incorrectly (rejected). This binary feedback can be provided immediately after the user speaks a sentence. In addition, aggregated scores can also be obtained by computing the average phone rejection rates at the sentence or speaker levels.

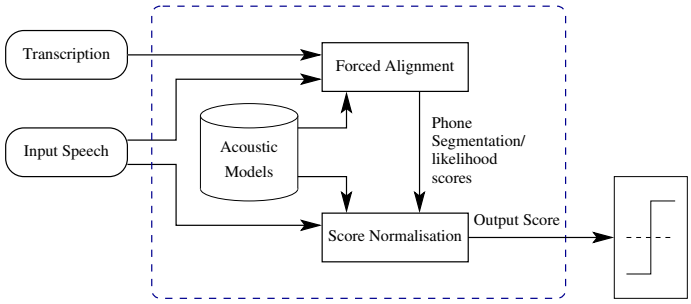


Fig. 1. System architecture of a typical pronunciation assessment system

The remaining of this paper is organised as follows. Section II gives a brief description of the proposed phone verification based pronunciation quality assessment system. Section III describes different goodness of pronunciation scores for phone verification. This is followed by the discussion of acoustic model training paradigm and feature normalisation techniques in Section **??** to improve phone verification performance. Finally, experimental results are presented in Section V.

II. System Description

This paper proposes a phone verification approach to automatic pronunciation quality assessment. This system generates

confidence scores at the phone level and makes a decision to accept or reject the pronunciations. The system architecture is illustrated in Figure 1. This system takes in an input speech and the corresponding orthographic transcription to produce output scores at the phone level. This approach can be viewed as an extension to the system introduced in [7]. The system can be divided into two major parts: confidence score generation and phone verification.

A. Confidence Score Generation

The first part of the score generation process is to perform a Viterbi forced aligment to obtain the phone segmentation, $S = \{s_i, e_i\}$, where s_i and e_i denote the start and end time of the ith phone segment respectively. The likelihood of the ith phone (denoted by m_i) given the segment boundaries ($\{s_i, e_i\}$) may be computed as

$$p(\mathcal{O}_i|m_i) = \sum_{Q_i} p(\mathcal{O}_i, Q_i|m_i) \approx \max_{Q_i} p(\mathcal{O}_i, Q|m_i) \qquad (1)$$

where $\mathcal{O}_i = \{o_{s_i}, \ldots, o_{e_i}\}$ and $Q_i = \{q_{s_i}, \ldots, q_{e_i}\}$ are the observation and state sequences for the ith segment. By assuming that the likelihood of the best path dominates, the summation operator in the above equation can be replaced by the $\max$ operator. This is the classic Viterbi likelihood approximation. The raw likelihood score produced by forced alignment needs to be normalised as confidence measures before they can be used for pronunciation verification. This confidence measure is also referred to as the Goodness of Pronunciation (GoP) [7]. Three types of GoP scores will be described in Section III.

B. Pronunciation Verification

By convention, the system output scores are produced such that a higher score indicates a higher confidence that the speech waveform aligned to a segment matches the corresponding phone. Therefore, a decision threshold can be applied to the confidence scores, above which the pronunciation is verified as being correct, and *vice versa*. The performance of a verification system is typically evaluated using the Equal Error Rate (EER) metric. EER is the average false acceptance and false rejection rates at the operating point where these errors are equal. EER provides a performance measure of a system under optimum operating condition.

In order to compute the EER for phone verification in our case, the reference phone sequence that the learner was supposed to utter is used to align the learner's speech to obtain the phone segmentation as described in Section II. Given the phone segmentation, the likelihood of a given model m generating the speech in a given phone segment can be computed using the Viterbi algorithm. Hence, if there were M phones in the system, there will be M scores for each phone segment; one of which is the *true* score while the remaining $M-1$ are *false* scores. By collecting all the true and false scores, the EER can be computed as described above. There are two ways of computing the overall EER of the system:

- **Pooled EER:** This approach pools all the true and false scores from all the phone segments and compute the EER based on a *global* decision threshold.
- **Average EER:** This approach computes the EER for each reference phone and then compute the overall EER by taking the average. This effectively applies a different decision threshold to different phones.

As presented in Section V, the average EER gives consistently better verification performance indicating that a different verification threshold should be applied to each type of sound.

In addition to evaluating the performance of a phone verification system, the EER computation also gives the optimum decision thresholds to be applied in the actual verification task. This provides a binary feedback at the phone level to the learner to indicate phone segments which the learner has or has not pronounced correctly. Moreover, it is also possible to provide an overall assessment score at the sentence and/or speaker level after the user has completed a learning lesson. The average phone rejection error rates can be aggregated to provide an overview of the error statistics. This allows the weakness of the user in pronouncing certain phones to be identified.

III. Goodness of Pronunciation

This paper examines three different methods of generating the GoP scores. These methods will be described in the following sub-sections.

A. Variable Segmentation Phone Posteriors

This method computes the phone posterior probability as:

$$P(Q_i|\mathcal{O}_i) = \frac{p(\mathcal{O}_i, Q_i)}{\sum_Q p(\mathcal{O}_i, Q)} \approx \frac{p(\mathcal{O}_i, Q_i)}{p(\mathcal{O}_i, Q_i^*)} \qquad (2)$$

where Q_i and Q_i^* correspond to the state sequence obtained by performing Viterbi forced alignment and phone-loop decoding respectively. The approximation in the above equation is made based on the assumption that the likelihood of the best state sequence dominates summation of the denominator. Since the segmentations of the two state sequences may be different, the denominator can be computed as

$$P(\mathcal{O}_i, Q_i^*) = \sum_q w_{qi} P(\mathcal{O}_q|q) \qquad (3)$$

where $\mathcal{O}_q$ denotes the observation which are aligned to state q and w_{qi} is the proportion of $\mathcal{O}_q$ which aligns with $\mathcal{O}_i$. This method is similar to the Goodness of Pronunciation (GoP) scores as proposed in [7] except that the average likelihood is computed at the state level instead of phone level.

B. Fixed Segmentation Phone Posteriors

If the phone segmentation is known, the phone posterior probability can be computed as:

$$p(m_i|\mathcal{O}_i) = \frac{p(\mathcal{O}_i|m_i)P(m_i)}{\sum_{m \in \mathcal{M}} p(\mathcal{O}_i|, m)P(m)} \qquad (4)$$

where $\mathcal{M}$ denotes the set of all phone models in the system. $P(m)$ denotes the probability of the phone m. This is typically

assumed to be a uniform distribution and can therefore be eliminated from the above equation. To compute the fixed segmentation phone posteriors, the phone segmentation can be obtained by performing a Viterbi forced-alignment.

C. Approximated Fixed Segmentation Phone Posteriors

The phone-level posterior probability in equation 4 can also be approximated by replacing the summation operator in the denominator with a max operator. This approximation tends to over-estimate the posterior probability which leads to slightly inferior results as presented in Section V.

IV. Constrained MLLR (CMLLR) Adaptation

Speaker adaptation is an important technique extensively used in speech recognition tasks to reduce variability due to different speakers. This paper investigates the use of Constrained MLLR (CMLLR) [2] speaker adaptation technique to improve phone verification performance. CMLLR speaker adaptation can be performed in two different modes: 1) supervised adaptation and 2) unsupervised adaptation. The former uses the true phone labels to estimate the adaptation parameters while the latter performs an initial unadapted phone recognition to obtain the phone labels. For language learning, the user's speech may contain mispronunciations. Hence, it is interesting to compare the performance of both adaptation modes.

V. Experimental Results

A. Experimental Setup

In this paper, the acoustic models used in the phone verification systems are trained on the Wall Street Journal (WSJ0) database recorded at Cambridge University [3]. It consists of 18.8 hours of training data collected from 92 different speakers. All acoustic models are trained using the 39 dimensional MFCC feature vectors, comprising 12 static coefficients, the `C0` coefficient plus the Δ and $\Delta\Delta$ differential parameters. Monophone HMM models were trained with a 3-state left-to-right topology and 32 Gaussian components per HMM state. HMM models were trained using HTK [8] Utterance-based cepstral mean normalisation was applied to reduce channel effects. To evaluate the phone verification performance, the evaluation data from the `MC-WSJ-AV` database [4] were used as test sets. Only the data collected using the `headset` and `lapel` were used to investigate the channel effects. Each set consists of 5 speakers with a total of 0.7 hours of data from 286 utterances.

B. Phone Verification Performance

TABLE I
Pooled and Average EER (%) of 32-component ML trained HMM models on the `headset` test set

GoP Score	EER (%)	
	Pooled	Average
VarSeg Posterior	8.39	8.04
Approx. FixSeg Posterior	6.02	5.93
FixSeg Posterior	5.92	5.79

This section compares the Pooled and Average EER performance for phone verification. Table I compares the performance of pooled *versus* average EER for phone verification using different goodness of pronunciation confidence scores. Clearly, using average EER leads to a consistent relative EER reduction of approximately 2.6%. This shows that it is useful to apply phone specific decision thresholds for phone verification. By comparing the performance of using different GoP scores, it was found that phone posteriors calculated using a variable phone segmentation yielded the worst average EER performance of 8.04%. By contrast, the fixed segmentation posterior scores gave significantly lower average EER performance with the approximated version being slightly inferior. This is due to the max assumption when computing the denominator term which leads to an over-estimate of the confidence scores. For subsequent experiments, the verification performance will be reported based on the average EER computed using the exact fixed segmentation phone posteriors.

TABLE II
Average EER (%) of 32-component HMM models on the `headset` and `lapel` test sets.

Test Set	Adaptation Mode	Average EER (%)		
		ML	MPE	MMI
headset	none	5.79	5.59	**5.21**
	unsup.	5.47	5.26	**5.00**
	sup.	5.23	5.07	**4.85**
lapel	none	10.52	**9.87**	9.92
	unsup.	9.88	**8.98**	9.04
	sup.	8.82	**7.87**	8.21

Next, the effects of acoustic model training paradigms and CMLLR adaptation using a global transformation were investigated. This paper compares two discriminative training methods: Maximum Mutual Information (MMMI) [1] and Minimum Phone Error (MPE) [6]. The average EER results are given in Table II. From this table, it was observed that the EER performance on the `lapel` test set is almost twice as high as those on the `headset` test set. Discriminative training methods consistently outperformed their ML counterpart. MMI was found to yield better results than MPE on `headset` set set. However, MPE was found to be better for the `lapel` test set.

After performing CMLLR adaptation, the EER performance consistently improved compared to the unadapted systems. For unsupervised adaptation, the MMI system improved by 4.0% and 8.9% respectively on the `headset` and `lapel` test sets respectively. With supervised adaptation, slightly better improvements can be achieved. The relative improvements over unadapted systems gave 6.9% and 17.2% respectively on the two test sets. As expected, CMLLR adaptation gave much larger improvements on the `lapel` test set due to channel mismatch between the training and testing conditions. Even after CMLLR adaptation, there is still a large gap in EER performance between the `headset` and the `lapel` test sets. This suggests that the sound quality captured using a lapel microphone is inherently poorer.

C. Score Aggregation Analyses

As previously mentioned in Section II-B, it is possible to provide an overall assessment score at the sentence and/or speaker level after the user has completed a learning lesson. For this purpose, speech data were collected from five graduate students with different country of origin. This paper only performs a preliminary comparison of the aggregated rejection rates at the speaker level. Table III shows the

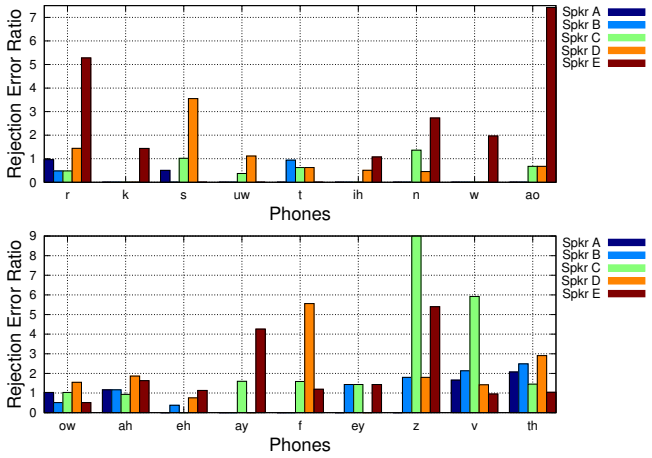
TABLE III
COMPARISON OF THE AGGREGATED REJECTION RATES OF FOUR DIFFERENT STUDENTS

Speaker	Country of Origin	First Language	Aggregated Rejection Rate (%)		
			none	unsup.	sup.
A	Singapore†	English	3.53	3.29	2.75
B	Indonesian	English	5.21	4.32	3.38
C	Germany	German	5.67	6.35	4.39
D	Singapore	English	10.02	7.15	4.79
E	China	Mandarin	8.90	7.68	5.32

† Spent most of the time in an English-speaking country

overall rejection rates for each speaker. These aggregated rejection rates were computed as the average aggregated rejection rates of each phone. Phone verification was performed using the 32-component MMI trained acoustic models. A global CMLLR transform was used for both supervised and unsupervised adaptation for each speaker. The speakers are ordered according the their fluency in spoken English. Without adaptation, the system has detected the order of Speaker D and E incorrectly. After adaptation, aggregated rejection rates successfully predicts the order of fluency of the five speakers. Note that the improvement after supervised adaptation is too optimistic that the rejection rates become similar to those of the native speakers (*c.f.* Table II). It may have corrected mispronunciations which is undesirable. Therefore, we feel that the unsupervised adaptation results provides a better assessment in this case. In addition, the breakdown

Fig. 2. Comparison of phone level aggregated rejection rates of four different speakers



of the phone level Rejection Error Ratios (RER) are shown in Figure 2. RER is the ratio of the rejection rates of the speaker divided by the rejection rates computed on a group of native speakers. It represents the number of times a speaker has made a pronunciation error compared to that made by the native speakers on average. Since the rejection rates are different for different phones, even when evaluated on native speech, RER provides a better comparison of rejection errors between different phones. Clearly, the rejection patterns differs significantly between speakers. For example, speaker A only made errors for six different phones, with '*th*' having the highest RER. On the other hand, speaker C has a relatively higher rejection rate for '*z*' and '*v*' while speaker E has difficulties pronouncing '*s*' and '*t*'.

VI. CONCLUSIONS

This paper has presented a phone verification approach to automatically assess the pronunciation quality at phone level for spoken language learning. The performance of the assessment system is evaluated using the equal error rate metric. The system provides a binary feedback of acceptance or rejection at the phone level. Aggregated rejection rates can also be computed at the phone, sentence and speaker levels to provide an overall assessment summary. In this paper, the following techniques were found to yield lower EER: phone-dependent decision thresholds, discriminative acoustic model training and constrained MLLR speaker adaptation. The EER performance of the best system were 4.85% and 7.87% on speech recorded using *headset* and *lapel* microphones respectively. Preliminary evaluation also revealed that the aggregated rejection rates were able to predict the learner's language competency level.

ACKNOWLEDGMENT

This research is done for CSIDM Project No. CSIDM-200806 partially funded by a grant from the National Research Foundation (NRF) administered by the Media Development Authority (MDA) of Singapore.

REFERENCES

[1] L. Bahl, P. Brown, P. deSouza, and L. Mercer. Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 49–52, 1986.

[2] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer. Speaker adaptation using constrained estimation of gaussian mixtures. *IEEE Transactions on Speech and Audio Processing*, 3:357–366, 1995.

[3] John Garofalo *et al.* CSR-I (WSJ0) complete. *Linguistic Data Consortium, Philadelphia*, 2007.

[4] M. Lincoln, I. McCowan, J. Vepa, and H.K. Maganti. The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): specification and initial experiments. *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop*, pages 357–362, 2005.

[5] Leonardo Neumeyer, Horacio Franco, Vassilios Digalakis, and Mitchel Weintraub. Automatic scoring of pronunciation quality. *Speech Communication*, 30:83–93, 2000.

[6] D. Povey and P. C. Woodland. Minimum Phone Error and I-smoothing for improved discriminative training. In *Proc.ICASSP*, 2002.

[7] S. Witt and S. Young. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30((2/3)):95–108, 2000.

[8] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. *The HTK Book (for HTK version 3.4)*. Cambridge University, December 2006.