DECISION FUSION FOR IMPROVING MISPRONUNCIATION DETECTION USING LANGUAGE TRANSFER KNOWLEDGE AND PHONEME-DEPENDENT PRONUNCIATION SCORING

W. K. LO[1,2], Alissa M. HARRISON[1], Helen MENG[1,2], Lan WANG[2]

[1]The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
[2]CAS-CUHK Shenzhen Institute of Advanced Integration Technologies, Shenzhen
{wklo, alissa, hmmeng}@se.cuhk.edu.hk, lan.wang@siat.ac.cn

ABSTRACT

Application of linguistic knowledge of language transfer to automatic speech recognition (ASR) technology can enhance mispronunciation detection performance in Computer-Aided Pronunciation Training (CAPT). This is achieved by pinpointing salient pronunciation errors made by second language learners. In this work, we propose to apply decision fusion for further improvement in mispronunciation detection performance. Detection decision from the linguistically-motivated detection, which applies language transfer knowledge, is used as the basis. Back off to posterior probability based pronunciation scoring with phoneme-dependent thresholds is employed when the basis is "less-reliable". Fusion can help combat problems such as incomplete coverage of linguistic knowledge as well as the imperfection of acoustic models in ASR. Our fusion strategy can maintain the diagnosis capability of the linguistically-motivated approach while achieve a major boost in detection performance. Experimental results show that decision fusion can achieve relative improvement in mispronunciation detection of up to 30% reduction in total number of decision errors.

Index Terms **—** pronunciation training, mispronunciation detection, pronunciation scoring, context-sensitive phonological rules, decision fusion

1. COMPUTER-AIDED PRONUNCIATION TRAINING

Computer-Aided Pronunciation Training (CAPT) is the application of spoken language technology to facilitate pronunciation training for language learners. It may include *production training* that teaches the learners how to pronounce as well as *perception training* that teaches learners how to distinguish different sounds of the language. In production training, CAPT usually performs automatic recognition on speech data collected from users and then makes decision on the correctness of pronunciation. There have been efforts in the investigation of automatic techniques to assign scores for collected speech data, e.g., [1-4]. Depending on the phonological characteristics of the target language and the desired granularity of analysis, scoring can be applied at the phone, syllable or word levels. By leveraging on automatic speech recognition (ASR) technology, a popular technique for pronunciation scoring is to compute the posterior probability of the speech unit in focus (e.g., a phone, a syllable, a word, etc.).

2. MISPRONUNCIATION DETECTION

2.1. Pronunciation Scoring by Posterior Probability

Pronunciation scoring is based on the posterior probability of the focused speech unit being produced by the speaker, given the acoustic observations and the speech recognizer (including model pronunciation, acoustic models, etc.) as shown in Eqn [1].

$$P\left(p|\,\vec{O},\,\Lambda\right)\approx\frac{ac\left(\vec{O}|p,\;\Lambda\right)}{\sum_{\forall p'\in P}ac\left(\vec{O}|p',\;\Lambda\right)} \quad [1]$$

where ac is the acoustic likelihood score, p is the focused speech unit, P is the set of all units (e.g. phoneme set), $\vec{O}$ is the acoustic observation, and Λ is the speech recognizer.

Depending on the needs of target users, this posterior probability can be used directly as the pronunciation score [1], or further categorized into grades [2] (e.g., normative scale from 1 to 5). Scoring pronunciation in this way can leverage existing ASR technologies to offer a quantitative (or categorical) assessment for the users. However, this family of techniques purely scores pronunciation by the acoustic signals without taking into consideration of the underlying linguistics. Incorporation of linguistic knowledge, especially those related to the salient errors of the target group of learners can hopefully improve the error detection performance.

2.2. Linguistically-motivated Pronunciation Error Detection

Our previous work [5, 6] showed that incorporation of phonological knowledge based on language transfer is useful for detecting salient pronunciation errors made by second language (L2) learners. This approach can detect errors and provide diagnostic feedback by specifying what kind of errors has been committed by the users. This diagnostic capability is brought about by the incorporation of language transfer considerations, which covers common errors committed by the learners due to the influence of their mother language (L1).

We performed mispronunciation detection [5, 6] as a recognition process where the model word pronunciations and possible erroneous word pronunciations (or variants) are all included in the recognizer's pronunciation lexicon. These possible erroneous pronunciations were obtained from the model pronunciations by the application of expansion rules derived from a contrastive phonological study between L1 and L2 of the L2 learners. The CAPT system detects mispronunciations by telling whether the pronunciation matches with the model pronunciation or if there may be error due to language transfer effects. In transcribing the learner's speech, the ASR needs to choose among the model pronunciations and the variants. In this way, the ASR can perform error detection and diagnosis at the same time.

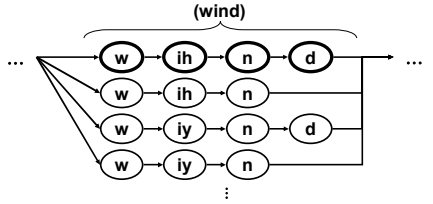

Figure 1. This example shows the expansion of the model pronunciation (thick border nodes) to pronunciation variants in an ASR pronunciation lexicon. The expansion is based on language transfer knowledge which captures possible errors frequently committed by L2 learners.

An illustrative example of the incorporation of language transfer knowledge in the form of an expanded ASR pronunciation lexicon is depicted in Figure 1. Schematic diagram of our system is also shown in Figure 2.

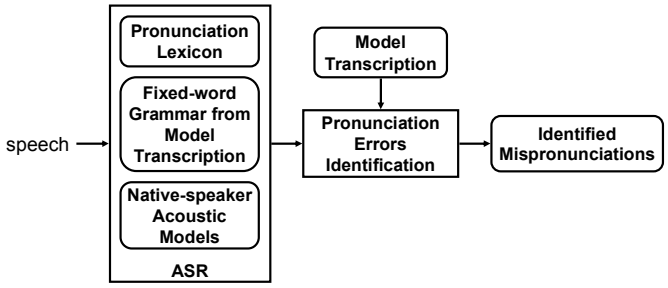

Figure 2. Schematic diagram showing the linguistically-motivated system for detecting and diagnosing L2 learners' mispronunciations.

2.3. Extension by Decision Fusion

The linguistically-motivated CAPT system described in the previous section focuses on detection of salient mispronunciations related to language transfer phenomena. Owing to the large number of possible expansions, pruning has been applied [5] or context dependency has been taken into consideration [6] for making this approach feasible. There are still cases where this approach may not perform well, especially for phonemes where: (i) the expansion rules are absent due to pruning or lack of relevant language transfer knowledge; (ii) the quality of the acoustic models is poor which in turn hinders recognition accuracy; (iii) the mispronunciations caused by factors outside the scope of our investigation (e.g., incorrect letter-to-sound conversion by the learners).

In this work, we aim to extend the previous approach for further improvement in mispronunciation detection performance by decision fusion. Since the linguistically-motivated approach possesses the diagnosis capability that pronunciation scoring does not have, we try to fuse the decisions by using the linguistically-motivated approach as a basis and *only* apply pronunciation scoring when the former approach is "less-reliable". The determination of "less-reliable" cases can be derived from some development data. With this knowledge, we can then take whichever the better decision among the component approaches. For example, back off can be applied for phonemes which do not have expansion rule or for phonemes whose acoustic models are poor in quality which hinder detection performance of the linguistically-based approach. Using pronunciation scoring in these situations can also take advantage of an additional degree of freedom by tuning the decision thresholds. The resulting fusion approach can combine the strength of the component approaches that the diagnosis capability is kept and at the same time a boost in mispronunciation detection performance can be achieved.

3. EXPERIMENTAL SETUP

3.1. Automatic Speech Recognition

The *acoustic models* used in this work are cross-word triphone HMMs. Every HMM has 3 states and every state has 12 Gaussian mixtures. The features adopted are 13-dimension PLP together with the first and second derivatives. Cepstral mean normalization is applied. There are altogether 6637 unique HMM states and 1987 unique models after state tying. The TIMIT corpus was used as our training data, which contains a total of 4620 sentences recorded by 462 speakers from eight dialect regions of the USA. The *recognition grammar* is generated dynamically according to the prompted reading materials read by the speakers. *Pronunciations* for each word in the reading materials are obtained from pronunciation dictionary and pronunciation variants are expanded according to the language transfer rules [5, 6].

3.2. Corpus

The test corpus used is a pilot collection of the CU CHinese Learners Of English (CU-CHLOE) corpus, which is the same corpus used in [6]. There are 21 speakers reading the Aesop fable "The North Wind and the Sun". This fable covers nearly all of the phonemes in English. There are only 4 (out of a total of 45) phonemes lack of data (/em/, /en/, /oy/ and /zh/). All data is manually transcribed by a linguist at the phoneme level. This test corpus is also divided into two disjoint subsets (SetA and SetB) in this work, while gender-balance is maintained. The subsets are used as development set (for parameter tuning) and test set *alternately* in order to verify the robustness of the investigated approaches. The speaker distribution is summarized in Table 1 together with the number of matched (correct) and mismatched (incorrect) pronunciations when compared to those obtained from a dictionary.

	# Male speakers	# Female speakers	# Correct	# Incorrect
Full Set	**9**	**12**	6600	1506
SetA	**5**	**6**	3451	795
SetB	**4**	**6**	3149	711

Table 1. Speaker distribution in the experimental sets. Speakers in SetA and SetB are mutually exclusive and their union is the Full Set. Some phonemes in SetA (hence in Full Set too) are mispronounced to the phonemes /en/ and /zh/ and therefore the no. of transcribed phonemes is more than 41.

3.3. Evaluation Criteria

In this work, we adopted the *total number of phoneme decision errors* (TotErr) for evaluation. This number is the sum of the number of false acceptance and the number of false rejection, at selected operating point. As with the usual convention in signal detection, the operating point is jointly determined by the false acceptance rate (FAR) and false rejection rate (FRR).

3.4. Baseline

Our previous work [6] shows that the linguistically-motivated detection approach can achieve additional performance improvement in mispronunciation detection by using context-dependent expansion rules. This approach is used as our baseline and the results on the test corpus are shown in Table 2.

Linguistically-motivated Approach			
	Full Set	**SetA**	**SetB**
TotErr	**1591**	**828**	**763**

Table 2. The detection performance of the linguistically-motivated approach using context-sensitive rules for pronunciation expansion.

4. DETECTION USING PRONUNCIATION SCORES

This section investigates the use of posterior probability based pronunciation scoring for mispronunciation detection at various operating points. This pronunciation score is basically follows Eqn [1] with minor modification. The time boundaries, *ts* and *te*, for every phoneme in the model transcription (a by-product from the linguistically-motivated approach) are used during computation of pronunciation score as shown in Eqn [2].

$$P\left(p|\bar{O}_{ts,\,te},\,\Lambda\right)\approx\frac{ac\left(\bar{O}_{ts,\,te}\mid p,\,\Lambda\right)}{\sum_{\forall p'\in P}ac\left(\bar{O}_{ts,\,te}\mid p',\,\Lambda\right)} \qquad [2]$$

where t_s is the start time, t_e is the end time and $\bar{O}_{t_s,t_e}$ is the segment of observation between t_s and t_e.

The numerator is given by the acoustic likelihood score of the model phoneme within the time boundaries (*ts* and *te*). For the denominator, we first compute the acoustic likelihood scores for each of the phonemes in the language (45 in total). The sum of these acoustic likelihood scores will then be used as the denominator. The overall procedures for computing the pronunciation score is summarized in Figure 3.

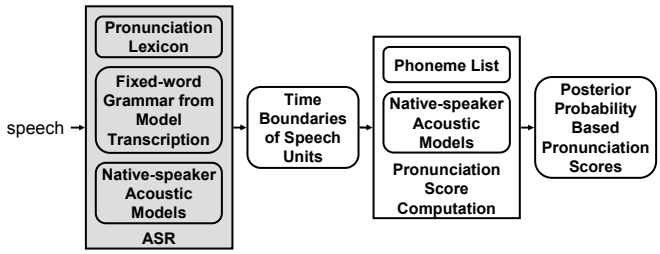

Figure 3. Schematic diagram shows the process of computing pronunciation scores. The grayed part is the same as that in the linguistically-motivated detection approach.

Figure 4 shows the detection results over the Full Set. Thresholding was used in the pronunciation scoring and we obtained the ROC at various operating points. More detailed analysis on these results is given in the following subsections.

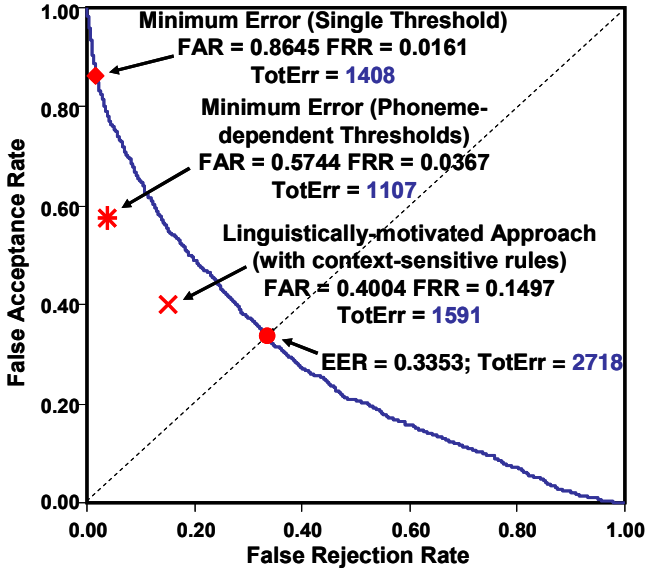

Figure 4. Detection results using pronunciation scoring based on posterior probability. "TotErr" refers to the total number of phoneme decision errors committed by the CAPT system. The "Diamond" is the operating point with minimum TotErr. The "Cross" is the operating point of the linguistically-motivated approach and the "Asterisk" is the minimum TotErr operating point with phoneme-dependent thresholds.

4.1. Detection with Single Threshold

When using pronunciation scoring, it can be observed that the mispronunciation detection performance is very poor (TotErr=2718) at the equal error rate (EER) operating point, when compared to the linguistically-motivated approach (TotErr=1591). Since the EER point is seldom used as operating point in a practical system, this result is shown for reference only.

A common operating point for a practical detection system is where TotErr is at its minimum. We can see from Figure 4 that at the minimum TotErr operating point (shown as a "Diamond"), a much better detection performance of just 1408 decision errors is achieved. The results on the Full Set and partitioned subsets are listed together in Table 3.

4.2. Phoneme-dependent Thresholding

When we looked into the detection performance for each of the phonemes individually, we found that the individual detection performance varies significantly. The optimal decision thresholds for individual phonemes also vary as can be seen in Figure 5. This can be attributed to the difference in complexity, quality and degree of mismatch with the learners' speech in the acoustic model for every phoneme. Figure 5 shows the variation of the optimal decision thresholds for minimum TotErr for each phoneme in the Full Set.

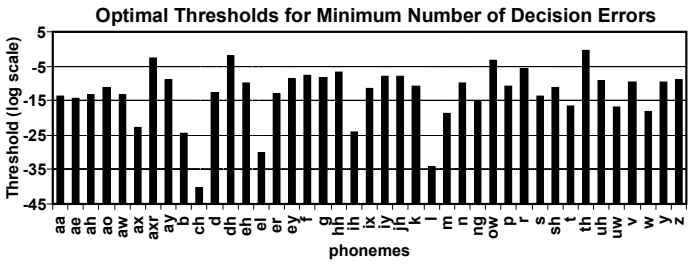

Figure 5. Variation of the optimal decision thresholds for minimum TotErr when carrying out detection over the Full Set using posterior probability based pronunciation scoring. The optimal single threshold is around -13.32 on the Full Set.

Based on this observation, we took a further step in improving the detection performance by using a phoneme-dependent decision threshold for each of the phonemes. The resultant optimal operating point for this approach is shown as an "Asterisk" in Figure 4. By using phoneme-dependent thresholds, an improved detection performance to just 1107 total decision errors is obtained. It is also the smallest number of errors among other approaches. The detection performance on the subsets SetA and SetB are shown in Table 3.

Detection Performance in TotErr using Pronunciation Scoring			
	Full Set	SetA	SetB
single threshold	1408	749	670
phoneme-dependent thresholds	1107	673	569

Table 3. The detection performance of the pronunciation scoring approach using a single threshold and using phoneme-dependent thresholds. All thresholds are optimized for minimum TotErr. It should be noted t hat the column for Full Set shows closed-set results for reference.

5. DECISION FUSION FOR MISPRONUNCIATION DETECTION

We have the linguistically-motivated approach which can both detect mispronunciations and diagnose the errors with the capability of diagnostic feedback. We also have an error detection-only approach using pronunciation scoring with phoneme-dependent thresholds. In order to combine the strength of these approaches, we used the linguistically-motivated approach as the basis (to keep its diagnosis capability) and backed off to pronunciation scoring with phoneme-dependent thresholds (for its mispronunciation detection performance).

5.1. Fusion procedure

We have investigated two fusion strategies for deciding when to back off. When we carried out fusion experiments, we made use of the partitioned sets as the development set and test set *alternately*, i.e., when tested on SetA, SetB is used as the development set, and *vice versa*. All decision thresholds and the list of phonemes to be backed off are derived from the development set results and applied to the test set.

Strategy A

The first fusion strategy is to back off to pronunciation scoring for phonemes which the linguistically-motivated approach makes consistent decision of accepting all instances without rejection in the development set. This is very likely an indication that there is a lack of expansion rule for these phonemes or that the content of the reading materials does not trigger any expansion rule. A list of such phonemes derived from the development set makes good candidates for back off.

Strategy B

The second strategy is purely performance based. We took the linguistically-motivated approach decision as the basis. When we encountered a phoneme where the pronunciation scoring achieved a *smaller TotErr* than the linguistically-motivated approach in the development set, we include such phoneme into the back off list. According to this list, we backed off to use pronunciation scoring during testing. This strategy not only takes into account of the available linguistic knowledge, it also considers other issues that may cause degradation in detection performance.

Table 4 summarizes the detection performance of the decision fusion and the relative improvement when compared to the component approaches.

Detection Performance in TotErr after Decision Fusion				
		Full Set	**SetA**	**SetB**
Baseline	Linguistically-motivated Approach	1591	828	763
	Pronunciation Scoring (phoneme-dependent thresholds)	1107	673	569
Decision Fusion	**Strategy A**	**1552**	**817**	**757**
	Strategy B	**1012**	**600**	**523**

Table 4. Detection performance in TotErr for the two fusion strategies. Relative improvements are w.r.t. the linguistically-motivated approach are 36.39%, 27.54% and 31.45% for the Full Set, SetA and SetB respectively. Similarly the relative improvements w.r.t. phoneme-dependent pronunciation scoring are 8.58%, 10.58% and 8.08% for the Full Set, SetA and SetB respectively.

6. DISCUSSION AND ANALYSIS

Fusion achieved improvement over component approaches

Our results show that decision fusion using the linguistically-motivated approach with back off to pronunciation scoring can bring improvement to mispronunciation detection. By using Strategy B (*back off when pronunciation scoring achieves a smaller TotErr)*, relative improvement over the linguistically-motivated approach is around 30% and that over pronunciation scoring with phoneme-dependent thresholds is around 9%, respectively.

It is found that Strategy A can only improve over the linguistically-motivated approach. Even though we found in our data that the pronunciation scoring detection results are superior for all of those backed off phonemes, the improved performance still cannot supersede that obtained from pronunciation scoring alone. This implies that there is room for a more aggressive fusion strategy (Strategy B) to back off when *pronunciation scoring achieves a smaller TotErr*.

Analysis of phonemes selected for back off

We compared the phonemes selected when using the Full Set, SetA and SetB as development set. There are several interesting points observed where the involved phonemes show consistent trends.

Cases when pronunciation scoring can help:

- Some phonemes are always (both strategies) backed off: /ay/, /ey/, /ix/, /jh/, and /w/. These phonemes include those with no expansion rules and at the same time pronunciation scoring detects pronunciation errors well.
- Some phonemes have rules and are not selected by Strategy A. They are backed off to pronunciation scoring by Strategy B: /ae/, /ao/, /b/, /er/, /g/, /ih/, /l/, /m/, /n/, /ng/, /r/, /uh/, /uw/, /v/, /y/, /z/. This implies that existing rules for these phonemes cannot predict the pronunciation errors well or the ASR performance is poor.

Cases when the linguistically-motivated basis is more reliable:

- Some phonemes *never* back off: /dh/, /ow/, /s/, /t/, and /th/. These include those with expansion rules and at the same time pronunciation scoring does not perform well.

For the remaining phonemes, they either do not show a consistent trend across the various partitions of the data set, or there is not sufficient data to have a conclusive result (e.g., phoneme /f/ in our data set are all pronounced correctly and hence we cannot evaluate the correct rejection and false acceptance of it). It should also be emphasized that by backing off to pronunciation scoring, the diagnosis capability of the linguistically-motivated approach is *lost*. For further extension to the current fusion strategies, we may opt for a back off only if the difference in TotErr between the component approaches is greater than some levels.

7. CONCLUSIONS

In our investigation for improving mispronunciation detection by decision fusion, we tested two strategies to fuse the linguistically-motivated approach and phoneme-dependent pronunciation scoring. It was found that significant improvement can be achieved by using the linguistically-motivated approach as a basis and back off to pronunciation scoring for phonemes that pronunciation scoring with phoneme-dependent thresholds commits fewer phoneme decision errors than the linguistically-motivated approach. This brings around 30% relative improvements over the linguistically-motivated approach in reduction of total phoneme decision errors. This indicates that the decision fusion strategy has successfully combined the strength of linguistically-motivated approach and pronunciation scoring for improving mispronunciation detection while maintaining the diagnostic capability of the linguistically-motivated approach.

8. ACKNOWLEDGMENTS

This work is affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies. It is also partially supported by the CUHK Teaching Development Grant.

9. REFERENCES

1 S. M. Witt and S. J. Young, "Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning," *Speech Communication*, vol. 30, pp. 95-108, 2000.

2 H. Franco, *et. al.*, "Combination of Machine Scores for Automatic Grading of Pronunciation Quality," *Speech Communication*, vol. 30, pp. 121-130, 2000.

3 G. Kawai and K. Hirose, "A Call System Using Speech Recognition to Teach the Pronunciation of Japanese Tokushuhaku," *STiLL1998*, pp. 73-76, Sweden, May 1998.

4 T. Kawahara, *et. al.*, "Practical Use of English Pronunciation System for Japanese Students in the CALL Classroom," *INTERSPEECH2004*, pp. 1689-1692, Korea, Oct 2004.

5 H. Meng, *et. al.*, "Deriving Salient Learners' Mispronunciations from Cross-Language Phonological Comparisons," *ASRU2007*, Japan, Dec 2007.

6 A. Harrison, *et. al.*, "Improving Mispronunciation Detection and Diagnosis of Learners' Speech with Context-Sensitive Phonological Rules Based on Language Transfer," *INTERSPEECH2008*, Australia, Sep 2008.