PERSONALIZED MISPRONUNCIATION DETECTION AND DIAGNOSIS BASED ON UNSUPERVISED ERROR PATTERN DISCOVERY

Ann Lee[1], *Nancy F. Chen*[2], *James Glass*[1]

[1]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts, USA
[2]Institute for Infocomm Research (I2R), A*STAR, Singapore
[1]{annlee, glass}@mit.edu, [2]nfychen@i2r.a-star.edu.sg

ABSTRACT

In this work, we introduce two improvements to our previously proposed mispronunciation detection framework. The framework focuses on each learner individually and consists of two main procedures: unsupervised error pattern discovery and pronunciation error decoding. First, we propose nbest filtering to disambiguate uncertain error candidate hypotheses obtained from acoustic similarity clustering. Second, we propose personalized template-based rescoring to refine the mispronunciation detection results. The second contribution of the paper is that we demonstrate the portability of the framework to a new target language. Experimental results on the iCALL corpus, a nonnative Mandarin corpus consisting of speakers of European origin, show that the new error pattern discovery process significantly reduces the size and increases the coverage of the error candidate set. Also, the rescoring technique effectively improves system performance on mispronunciation detection and diagnosis.

***Index Terms*—** Computer-assisted pronunciation training (CAPT), dynamic time warping (DTW), extended recognition network (ERN)

1. INTRODUCTION

There has been rapid growth in the number of people with various native language (L1) backgrounds learning a second language (L2). With the pronunciation assessment and corrective feedback provided by computer-aided pronunciation training (CAPT) systems, students enjoy great flexibility in both time and place when practicing their speaking skills [1, 2]. According to language transfer theory, learners often apply knowledge from their L1 to an L2 [3], which indicates that learners coming from the same L1 background will very likely share the same set of pronunciation error patterns. Nevertheless, other factors, such as a learner's level of competency in the L2, also affect error patterns [1, 4]. As a result, students often learn better with individual tutoring than with conventional classroom teaching.

However, the concept of personalization in current CAPT systems exists only in forms such as student performance tracking [5] and automatic material generation [6]. Most existing mispronunciation detection algorithms rely on a large amount of labeled training data and linguistic knowledge in the L1s and L2s [7, 8, 9, 10, 11]. As a result, the system's assessment ability is constrained by the available resources, and it is difficult to tailor a system to every student.

In our previous work [12], we proposed a mispronunciation detection system that does not require nonnative training data. The system contains two main procedures: unsupervised error pattern discovery and pronunciation error decoding. In the first stage, we discover an individual learner's error patterns by computing acoustic distance between phoneme segments. In the second stage, on the basis of the learner-specific error candidates and the context constraint that only one type of pronunciation error is allowed per triphone, we decode phonemic pronunciation errors produced by Cantonese and Mandarin (L1) learners of English (L2).

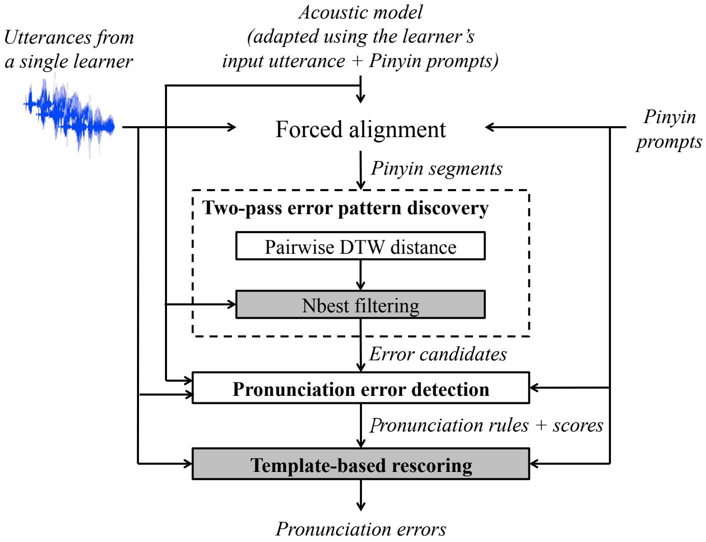

Fig. 1. System flowchart. The system focuses on a single learner and performs two-pass error pattern discovery to determine a set of error candidates. Pronunciation error detection and template-based rescoring are then carried out to determine the actual pronunciation errors. Shaded blocks are newly proposed in this work (discussed in Section 3.1 and 3.3).

In this work, we demonstrate the portability of the proposed framework by focusing on phonemic pronunciation errors produced by English (L1) learners of Mandarin (L2). Due to the emergence of mobile-assisted language learning apps [13], collecting speech from a single learner has become easy [14]. Therefore, we expand the scale of experiments by testing the framework using the iCALL corpus [15, 16], a large-scale nonnative Mandarin corpus consisting of speakers of European origin. In addition, the framework is improved in two directions. As shown in Fig. 1, for the first stage, we run nbest filtering after computing acoustic distances to form a two-pass error pattern discovery process. Experimental results show that the new error candidate set has higher quality in terms of error coverage and size. Second, we perform personalized template-based rescoring using a set of segments selected from the learner's speech. Experimental results show that the likelihood scores from the recognizer and the distance scores from the templates are complementary.

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

2. RELATED WORK AND BACKGROUND

2.1. Mispronunciation detection and diagnosis

Automatic speech recognition (ASR) technology has been a core component in CAPT systems [1, 2]. Distance scores from template-based recognizers [17] and likelihood and posterior probability scores from hidden Markov model (HMM)-based recognizers [18, 19] have been computed to detect pronunciation errors. In order to diagnose the exact error types, some prior work identifies error patterns either from expert knowledge [7, 9, 10, 11] or from nonnative training data [7, 8], and incorporates them into the lexicon to form an extended recognition network (ERN). During decoding, the errors and the error types are detected at the same time.

However, both the linguistic expertise and a fully transcribed nonnative corpus are expensive and time-consuming to collect. Molina et. al [20] propose to generate possible confusion words based on distance between acoustic models from an ASR engine. Wang and Lee [21] perform unsupervised phoneme posteriorgram clustering to discover mispronunciation patterns directly from data. Our proposed framework in [12] is based on ERNs. Instead of extracting error patterns from nonnative training data, we compute acoustic distance between phoneme segments from an individual learner's speech to determine a set of learner-specific error patterns in an unsupervised manner.

2.2. Mandarin Chinese phonology

Each character in Mandarin Chinese corresponds to a single syllable. A syllable starts with an optional initial (consonant), and then a final (vowel and an optional nasal consonant) together with a lexical tone.

In this work, we focus on phonemic pronunciation errors and leave tone errors aside for now. We refer to initials and finals as *Pinyin units*, and they are the target units for our mispronunciation detection task. In addition, similar to the concept of triphone in English speech recognition, we use a concept of *tripinyin*, which consists of a Pinyin unit together with its left and right contexts. For instance, a tripinyin *#_zh_ong* is an initial *zh* under the context of a final *ong*, and # stands for the syllable boundary.

2.3. Extended recognition network (ERN)

In a finite state transducer (FST) based recognizer, the lexicon is represented as an FST that maps phoneme sequences to words. The FST can be enhanced by adding multiple arcs corresponding to possible phoneme variations, and thus form an ERN. For Mandarin Chinese, we incorporate the error patterns on the Pinyin unit level. Fig. 2 shows an example of an ERN of the character (in Pinyin) "*zhong*", with one deletion on the initial and one substitution on the final. Running recognition or forced alignment with the ERN may result in output Pinyin sequences different from the canonical pronunciations, and thus we can detect pronunciation errors.

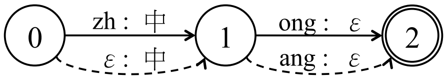

Fig. 2. An example of an ERN of the character (in Pinyin) "*zhong*". Error patterns are incorporated on the Pinyin unit level (initial and final), as shown in dashed arcs. ϵ denotes an empty string in FST input/output. In this work we do not consider tone errors.

3. SYSTEM DESIGN

We discuss the new features of the framework as shown in Fig. 1.

3.1. Two-pass unsupervised error pattern discovery

After forced alignment with canonical pronunciations, assume the utterances are segmented into N Pinyin units, $\{s_i\}_{i=1}^N$. Let $\mathbf{U} = \{u_i\}_{i=1}^U$ be the Pinyin unit inventory set, and $\mathbf{R} = \{r_i\}_{i=1}^R$ be the set of unique tripinyin patterns. Each segment s_i is associated with one canonical Pinyin label $p_i \in \mathbf{U}$ and one tripinyin label $t_i \in \mathbf{R}$.

In the first pass, we compute dynamic time warping (DTW) distance between frames of MFCCs for all (s_i, s_j) pairs, denoted $DTW(s_i, s_j)$. A threshold τ is set so that we obtain a set of Pinyin labels for each s_i from their nearest neighbors as *local candidates*:

$$C_L^i = \{p_j | p_j \neq p_i, DTW(s_i, s_j) \leq \tau\}, i = 1, ..., N \qquad (1)$$

We gather the candidates from segments with the same tripinyin label and form R first-pass tripinyin-specific error candidate sets:

$$EC^i_{first} = \bigcup_{j, t_j = r_i} C_L^j, i = 1, ..., R \qquad (2)$$

Note that we do not consider global candidates obtained based on Gaussian mixture model (GMM) distance as in previous work [12], as a pilot study showed that using a DTW distance generates more accurate error candidate sets than a GMM distance does.

As EC_{first}'s are obtained based on acoustic distances between segments, the direction of mispronunciation, i.e. whether p_i is mispronounced as p_j or the opposite, is unclear. To disambiguate this uncertainty, we propose to run a second pass of nbest filtering. For each EC^i_{first}, we build an ERN by incorporating the error candidate set into the canonical lexicon. We run R times of forced alignment using one ERN at a time. In the end, EC^i_{second} consists of the Pinyin labels from the i-th time nbest output.

3.2. Pronunciation error decoding

We convert the R error candidate sets into L pronunciation rules by considering substitution and deletion errors only, as insertions are rare based on empirical analysis. A deletion error is considered for an initial when there is a final in its candidate set. No deletion is allowed for the finals as they are the nucleus of a syllable, and we assume the learners read all the Pinyin symbols.

We run decoding with the constraint that only one pronunciation rule is allowed per tripinyin as in previous work. In the iterative decoding process proposed in [12], the first iteration runs L times of forced alignment, each time with an ERN that incorporates only one pronunciation rule. The rule with the best likelihood score is chosen and incorporated into the lexicon, and the process continues on to the next iteration of forced alignments with the remaining rules. To deal with the large scale of speech that might be emerging in the future, in this work, we approximate the iterative process by running the first iteration (L times of forced alignment) only and select rules based on their likelihood scores. This approximation improves the worst case time complexity from exponential to linear, while a pilot study did not show significant degradation in system performance.

3.3. Speaker-dependent template-based rescoring

In CAPT, an inevitable challenge is the mismatch between the acoustic characteristics of native and nonnative speech. In our personalized framework, we tackle this challenge by taking advantage of the large amounts of per-speaker data to build a speaker-dependent template-based speech recognizer. It has been empirically shown that template-based approaches complement standard parametric ASR systems in speech retrieval tasks [22].

The templates consist of the segments that are marked as correct from all the L times of forced alignment in the previous step. Two types of templates are built. The first type contains templates for each tripinyin r_i:

$$M^i_{tri} = \{s_j | t_j = r_i, s_j \text{ is correct}\}, i = 1, ..., R \tag{3}$$

The second type contains templates for each unique Pinyin unit u_i:

$$M^i_{mono} = \{s_j | p_j = u_i, s_j \text{ is correct}\}, i = 1, ..., U \tag{4}$$

To compute the distance score for a pronunciation rule l_i, we first locate the mispronounced segments from the forced alignment result. Assume the format of l_i is α_β_γ (tripinyin) $\rightarrow$ δ (Pinyin unit). The distance score of l_i is computed as follows.

1. If δ represents a deletion error, since we assume finals cannot be deleted, β can only be an initial and γ a final, and the target tripinyin is $\#\_\gamma\_\#$. Otherwise, δ is a substitution, and the target tripinyin is α_δ_γ.
2. We compute average pairwise DTW distance between the mispronounced segments and the templates from the set M_{tri} of the target tripinyin ($\#\_\gamma\_\#$ or α_δ_γ). If the template set is empty, we back off to use M_{mono} of δ (or M_{mono} of γ if δ is deletion).
3. Average DTW distance with M_{tri} of α_β_γ (or back off to M_{mono} of β) is computed as reference distance.

The distance score is defined as the distance to the target template set minus the reference distance. It can be viewed as a confidence score from the learner-specific template-based speech recognizer. The final score for each rule l_i is a weighted sum between the negative log likelihood score from the forced alignment and the distance score.

4. EXPERIMENTS

4.1. Corpus

The nonnative corpus used in this study is the iCALL corpus [15, 16]. It consists of 305 beginning learners of Mandarin Chinese from European origin reading 300 Pinyin prompts, including 200 words and 100 sentences. All utterances are manually transcribed in Pinyin. The difference between the transcription and the Pinyin prompts serves as the ground truth pronunciation errors for evaluation. In addition, every utterance has a proficiency score graded by an expert, ranging from 1 to 4, with 4 being the highest level.

Table 1 shows the division of the data for experiments. In our current study, we only consider learners with English as L1 to avoid L1 mismatch issue for the supervised oracles. Compared with our previous study, new experiments are on a much larger scale (from 200 triphones per speaker to 500 tripinyins per speaker).

Table 1. Division of the corpus for experiments. Only speakers whose L1 is English are considered.

Speakers	# Pinyin units	Average proficiency score
Training (for oracles)		
36 males, 25 females	148,592	2.63
Testing		
34 males, 22 females	134,637	2.72

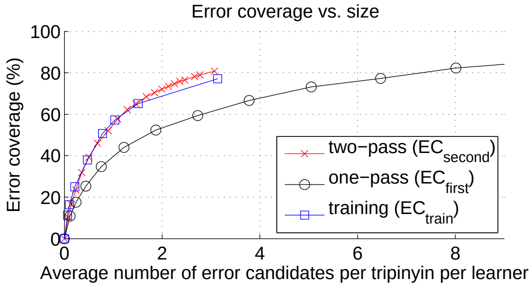

Fig. 3. Evaluation of the quality (error coverage vs. size) of the error candidate sets. The nbest filtering process reduces the size of EC_{first} by more than half.

4.2. Experimental settings

The native acoustic model is trained using the GALE phase 2 Mandarin Chinese broadcast news corpus with the Kaldi toolkit [23]. A 120-hr subset is randomly chosen as the training set, and the remaining 6-hr as the development set. All waveforms are transformed into 39-dimensional MFCCs plus three-dimensional pitch features every 10-ms, including first and second order derivatives. Cepstral mean normalization (CMN) is done on a per speaker basis, followed by linear discriminant analysis (LDA) and feature-space maximum likelihood linear regression (fMLLR) for feature transformation. Each Pinyin unit is mapped to a unique phoneme sequence, and we build a subspace GMM (SGMM)-HMM-based triphone model trained with maximum mutual information (MMI) criterion [24]. The character error rate on the development set is 13.26%.

Speaker adaptation is done for each learner separately. As students make less mistakes on shorter utterances [16], only utterances whose length is less than three characters and their canonical pronunciation from the Pinyin prompts are included into MMI training.

4.3. Error pattern coverage

We first examine the error candidate sets generated under three scenarios: two-pass error pattern discovery (EC_{second}), one-pass error pattern discovery (EC_{first}), and error patterns extracted from the training data (EC_{train}). We evaluate their quality by computing their size and error coverage, which is the percentage of error patterns in the ground truth that are found in the error candidate set.

Fig. 3 shows the average number of error candidates per tripinyin versus the error coverage of the test set. As the DTW threshold τ and the length of the nbest list increase, the size and the coverage of EC_{first} and EC_{second} also increase. As τ approaches infinity, EC_{first} becomes the Pinyin inventory and its coverage reaches 100%. On the other hand, a threshold can be set on the frequency of occurrences of the error patterns in the training set. Both the coverage and the size of EC_{train} decrease as the frequency occurrence threshold increases, and there exists limitation on the maximum coverage for EC_{train}, which is 77% in this case. The nbest filtering removes redundancy in EC_{first}. When the average error coverage is at 77%, the size of EC_{second} is 59% smaller than the size of EC_{first}. A compact error candidate set makes the subsequent pronunciation error decoding process more efficient.

Although the trends in size versus error coverage from EC_{second} and EC_{train} are similar, in Fig. 4, we further examine their worst case coverage, i.e. the minimum coverage on an individual learner.

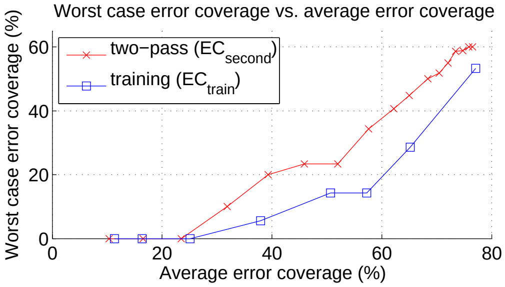

Fig. 4. Evaluation of the minimum error coverage on a single learner. EC_{second} consistently has higher worst case coverage, implying stronger tolerance in speaker variation.

The result shows that EC_{second} consistently has higher worst case coverage than EC_{train}. This demonstrates the advantage of performing the error candidate selection process on a per learner basis. The error patterns in EC_{train} may represent the common problem of an L1 population, but it may not cover individual variation.

4.4. Mispronunciation detection and diagnosis

We evaluate the system's final performance using EC_{second} (unsupervised, *unsup*) and EC_{train} (supervised, *sup*), with the operating points set at where the error coverage is 77%. The weight on the distance score for rescoring is empirically set to 60. Results based on EC_{train} are considered as *oracles* as they use information from the training data. An unsupervised baseline is implemented based on free Pinyin unit recognition (*pinyin-rec*) with a trigram Pinyin unit language model also trained on the GALE Mandarin Chinese corpus. Pronunciation errors are detected when there is mismatch between the *pinyin-rec* output and the Pinyin prompts.

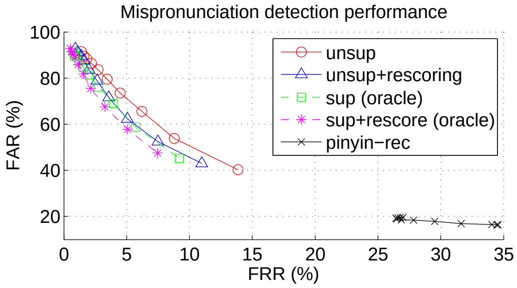

Fig. 5. Evaluation of false acceptance rate (FAR) vs. false rejection rate (FRR). Speaker-dependent template-based rescoring improves the system's performance by reducing both FAR and FRR.

4.4.1. Mispronunciation detection

We compute two metrics to evaluate the performance of mispronunciation detection: false rejection rate (FRR), which is the percentage of the total number of correct Pinyin units that are misidentified by the system, and false acceptance rate (FAR), which is the percentage of the number of all the incorrect Pinyin units that are accepted by the system as correct.

Fig. 5 shows the results. A threshold can be set on the scores of the pronunciation rules to control the size of system's output, which leads to trade-off between FRR and FAR. We also adjust the language model weight in free Pinyin unit recognition. However, it constantly has high FRR ($> 25\%$), and this is unfavorable for a CAPT system as it would discourage the students [1]. As a result, we focus on the remaining four system settings in the following discussion and evaluation.

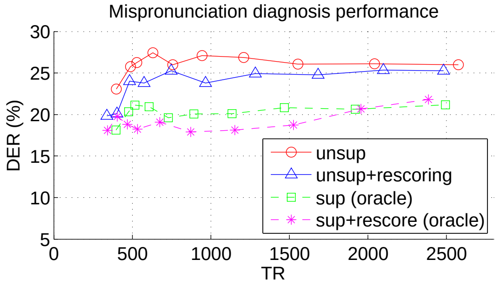

Fig. 6. Evaluation of diagnostic error rate (DER) vs. number of true rejection (TR). Speaker-dependent template-based rescoring also helps reduce DER of the proposed framework.

Speaker-dependent template-based rescoring improves the performance for both the *unsup* and *sup* systems. When FAR is at 50%, FRR of the *unsup* system is improved by 14% relative, and 11% relative for the *sup* systems. With templates from the learner's own speech, speaker variation is removed and thus it helps distinguish between acoustically similar segments on which the recognizer often fails. In fact, the performance of *unsup+rescore* achieves the same performance as the oracle system without rescoring.

4.4.2. Mispronunciation diagnosis

We focus on the segments whose ground truth errors are covered by the error candidate sets and compute diagnostic error rate (DER), which is the percentage of the correctly detected pronunciation errors (true rejection, TR) that have incorrect diagnostic feedback.

Fig. 6 shows the results. In high TR region, the performance of DER is more steady. When TR is at 1500, template-based rescoring helps reduce DER by 5% relative for the *unsup* system. There remains a gap towards the oracles. One explanation is that the templates contain segments which are mispronounced and thus introduce noise in distance scores. Also, the error candidates in EC_{second} are more acoustically similar to each other than the candidates in EC_{train} due to the nature of the unsupervised error pattern discovery process.

5. CONCLUSION AND FUTURE WORK

In this paper, we present two improvements to our previously propose mispronunciation detection system and demonstrate the L2 language portability of the framework by empirical validation on nonnative Mandarin. Experimental results show that the proposed nbest filtering reduces the size of the error candidate set by 59% relative while preserving its complete coverage. Furthermore, we are able to accommodate to high variations of error patterns across learners and thus achieve better worst-case error coverage. In addition, the proposed personalized template-based rescoring effectively reduces detection errors in an unsupervised fashion.

We have been focusing on the lexical level in ASR by modeling mispronunciations with ERNs. Therefore, while the automatically discovered error candidate set has high error coverage, the performance is limited by the acoustic model. As the community has started to explore deep learning in mispronunciation detection [25, 26], we also plan to improve performance from the acoustic level.

6. ACKNOWLEDGEMENT

The authors would like to thank Tong Rong and Boon Pang Lim from I2R for their help in setting up ASR experiments and in improving our framework.

7. REFERENCES

[1] M. Eskenazi, “An overview of spoken language technology for education,” *Speech Communication*, 2009.

[2] S. M. Witt, “Automatic error detection in pronunciation training: where we are and where we need to go,” in *Proc. IS ADEPT*, 2012.

[3] R. Ellis, *Classroom second language development: A study of classroom interaction and language acquisition*, Oxford University Press, 1984.

[4] L.F. Bachman, *Fundamental considerations in language testing. Oxford University Press*, Oxford University Press, 1990.

[5] “EnglishCentral,” http://www.englishcentral.com/.

[6] P.-H. Su, Y.-B. Wang, T.-H. Yu, and L.-S. Lee, “A dialogue game framework with personalized training using reinforcement learning for computer-assisted language learning,” in *Proc. ICASSP*, 2013.

[7] C. Cucchiarini, H. Van den Heuvel, E. Sanders, and H. Strik, “Error selection for ASR-based English pronunciation training in “my pronunciation coach”,” in *Proc. Interspeech*, 2011.

[8] W. K. Lo, S. Zhang, and H. Meng, “Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system,” in *Proc. Interspeech*, 2010.

[9] J. Kim, C. Wang, M. Peabody, and S. Seneff, “An interactive English pronunciation dictionary for Korean learners,” in *Proc. Interspeech*, 2004.

[10] H. Meng, Y.Y. Lo, L. Wang, and W.Y. Lau, “Deriving salient learners’ mispronunciations from cross-language phonological comparisons,” in *Proc. ASRU*, 2007.

[11] A. M. Harrison, W. Y. Lau, H. Meng, and L. Wang, “Improving mispronunciation detection and diagnosis of learners’ speech with context-sensitive phonological rules based on language transfer,” in *Proc. Interspeech*, 2008.

[12] A. Lee and J. Glass, “Mispronunciation detection without nonnative training data,” in *Proc. Interspeech*, 2015.

[13] R. Godwin-Jones, “Emerging technologies: mobile apps for language learning,” *Language Learning & Technology*, 2011.

[14] C.-F. Yeh, H.-Y. Lee, and L.-S. Lee, “Speaking rate normalization with lattice-based context-dependent phoneme duration modeling for personalized speech recognizers on mobile devices.,” in *Proc. Interspeech*, 2013.

[15] N. F. Chen, V. Shivakumar, M. Harikumar, B. Ma, and H. Li, “Large-scale characterization of Mandarin pronunciation errors made by native speakers of European languages,” in *Proc. Interspeech*, 2013.

[16] N. F. Chen, R. Tong, D. Wee, P. Lee, B. Ma, and H. Li, “iCALL corpus: Mandarin Chinese spoken by non-native speakers of European descent,” in *Proc. Interspeech*, 2015.

[17] D. Kewley-Port, C. Watson, D. Maki, and D. Reed, “Speaker-dependent speech recognition as the basis for a speech training aid,” in *Proc. ICASSP*, 1987.

[18] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, “Automatic detection of phone-level mispronunciation for language learning,” in *Proc. Eurospeech*, 1999.

[19] S.M. Witt and S.J. Young, “Phone-level pronunciation scoring and assessment for interactive language learning,” *Speech Communication*, 2000.

[20] C. Molina, N. B. Yoma, J. Wuth, and H. Vivanco, “ASR based pronunciation evaluation with automatically generated competing vocabulary and classifier fusion,” *Speech communication*, 2009.

[21] Y.-B. Wang and L.-S. Lee, “Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning,” *IEEE/ACM Transactions on Audio, Speech & Language Processing*, 2015.

[22] H. Xu, P. Yang, X. Xiao, L. Xie, C.-C. Leung, H. Chen, J. Yu, H. Lv, L. Wang, S. J. Leow, B. Ma, E. S. Chng, and H. Li, “Language independent query-by-example spoken term detection using n-best phone sequences and partial matching,” in *Proc. ICASSP*, 2015.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi speech recognition toolkit,” in *Proc. ASRU*, 2011.

[24] R. Tong, B. P. Lim, N. F. Chen, and H. Li, “Subspace Gaussian mixture model for computer-assisted language learning,” in *Proc. ICASSP*, 2014.

[25] W. Hu, Y. Qian, and F. K. Soong, “An improved DNN-based approach to mispronunciation detection and diagnosis of L2 learners speech,” in *Proc. SLaTE*, 2015.

[26] K. Li, X. Qian, S. Kang, P. Liu, and H. Meng, “Integrating acoustic and state-transition models for free phone recognition in L2 English speech using multi-distribution deep neural networks,” in *Proc. SLaTE*, 2015.