IMPROVING NON-NATIVE MISPRONUNCIATION DETECTION AND ENRICHING DIAGNOSTIC FEEDBACK WITH DNN-BASED SPEECH ATTRIBUTE MODELING

Wei Li[1], Sabato Marco Siniscalchi[1,2], Nancy F. Chen[3], and Chin-Hui Lee[1]

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA. USA
[2]Department of Telematics, Kore University of Enna, Enna, Italy
[3]Institute for Infocomm Research, Singapore
lee.wei@gatech.edu, marco.siniscalchi@unikore.it, nfychen@i2r.a-star.edu.sg, chl@ece.gatech.edu

ABSTRACT

We propose the use of speech attributes, such as voicing and aspiration, to address two key research issues in computer assisted pronunciation training (CAPT) for L2 learners, namely detecting mispronunciation and providing diagnostic feedback. To improve the performance we focus on mispronunciations occurred at the segmental and sub-segmental levels. In this study, speech attributes scores are first used to measure the pronunciation quality at a sub-segmental level, such as manner and place of articulation. These speech attribute scores are integrated by neural network classifiers to generate segmental pronunciation scores. Compared with the conventional phone-based GOP (Goodness of Pronunciation) system we implement with our dataset, the proposed framework reduces the equal error rate by 8.78% relative. Moreover, it attains comparable results to phone-based classifier approach to mispronunciation detection while providing comprehensive feedback, including segmental and sub-segmental diagnostic information, to help L2 learners.

Index Terms — mispronunciation detection and diagnosis, automatic speech recognition (ASR), deep neural network (DNN), computer assisted pronunciation training (CAPT), automatic speech attribute transcription (ASAT)

1. INTRODUCTION

With accelerating globalization, more and more people are willing or required to learn a second language (L2), in addition to their mother tongue language (L1) [1, 2]. The shortage of qualified L2 language teachers has thus become a serious problem, and computer assisted language learning (CALL) systems can play a key role in alleviating the lack of qualified teachers. An essential component of CALL is computer assisted pronunciation training (CAPT) used to automatically detect learners' mispronunciations and ideally provide diagnostic feedbacks. Mispronunciations refer to pronunciation errors, where surface pronunciation forms differ from correct canonical pronunciation forms. These errors can be defined on various time-scales, namely supra-segmental (e.g., lexical stress, intonation, lexical tones etc.), segmental (e.g., substitution of phonetic units), and sub-segmental (e.g., voicing feature activated for a canonical unvoiced phone [3, 4])**.**

Over the past decade, automatic speech recognition (ASR) systems have been employed to assess the goodness of pronunciation at the segmental level. For example, Log-Likelihood Ratio (LLR) between native-like and non-native models was employed in [5] to detect mispronunciation errors. Witt & Young [6] introduced "Goodness of Pronunciation" (GOP)", which is a generalized log likelihood ratio score, known to be used in utterance verification [7-11], taking into account the likelihood of both the intended canonical phone and a pool of competing phones. Subsequently, several variants of the GOP score were proposed [12, 13, 14]. Although good mispronunciation detection results can be attained using the above-mentioned systems, the feedback at the segmental level might not be intuitive or instructive enough for the L2 learner to correct his mispronunciation.

Another CAPT framework, called extended recognition network (ERN), had also been proposed [15, 16, 17]. The phone recognition network is first expanded by adding common phonetic error patterns. Then this ERN is used to force align learners' utterances. By contrasting the canonical form with the forced-aligned spoken utterances, the ERN method can provide some diagnostic information related to phone substitution, i.e., phone /A/ has been substituted with phone /B/. Nonetheless, a major assumption made by providing a feedback at a segment (a.k.a. phone) level is that learners are aware of which articulatory movements (e.g., manner and place of articulation [3, 4]) have to be corrected in order to restore the canonical phone pronunciation. Unfortunately, that is a challenging task for L2 beginners. For example, some phones in L2 are absent in L1: beginner Japanese learners of English might pronounce "lice" instead of "rice". Facing the segmental level feedback "phone r is mispronounced as phone l", learners might fail to adjust their articulatory movements to correct this error, because the phone /r/ does not exist in Japanese. Moreover, even if the supposed target phone exists in the learner's L1, the acoustic realizations of it might differ from the target language. For example, in Indian English and Singapore English, dental fricatives might be pronounced with stop-like acoustic features, making the word "three" sound like "tree" [18].

Exploiting information at a sub-segmental level enhances the feedback quality and alleviates some of the problems mentioned above. Facing the same mispronunciation described above, the Japanese learners could be instructed to improve their pronunciation if they were given sub-segmental feedback "make a sound similar to /l/ but roll your tongue more backwards to create the acoustic characteristics of /r/". Indeed, it has been reported that L2 learners prefer to receive direct instruction on how to correct mispronunciation at a sub-segmental level [19]. Moreover, mispronunciation detection at a sub-segmental level can more accurately specify systematic L2 pronunciation errors [20]. Attracted by such potential benefits, researchers have exploited sub-segmental information for L2 learning [19, 21, 22], where an acoustic-to-articulatory inversion method is adopted to directly

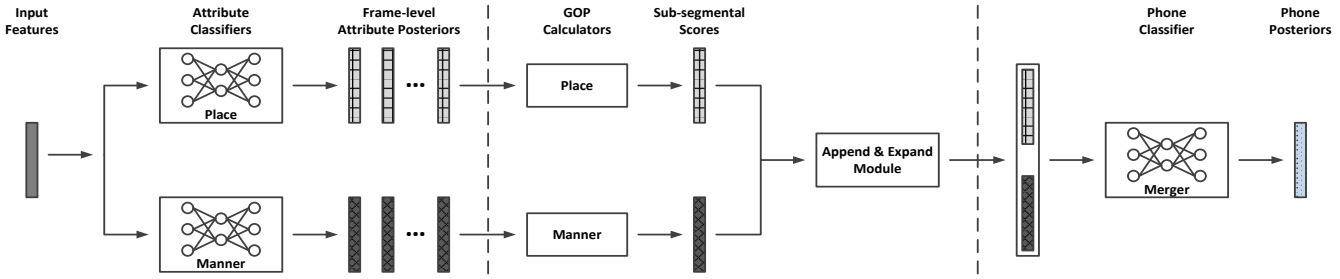

Figure 1: *Overview of mispronunciation detection framework, adapted from [27].*

provide feedback at an articulatory level. In [20, 23, 24], rule-based acoustic-articulatory mapping tables were employed to overcome the difficulty of collecting articulatory measurements to map each phone to its corresponding articulators. However, past work in mispronunciation detection performance at the sub-segmental level have been suboptimal due to the use of shallow models, or the lack of large training corpora [20].

Recently, iCALL, a large non-native speech corpus, has been designed, collected, and annotated at the Institute for Infocomm Research (I2R) [25]. A large training corpus can facilitate the full usage of deep neural networks (DNNs) for better sub-segmental estimation and detection [26, 27, 28]. In this work, *speech attributes*, such as voicing, aspiration, and manner and place of articulation, are extracted with DNNs. We use the terms speech attributes, sub-segmental, and acoustic-phonetic interchangeably in this work. The frame-based DNN posteriors are used as scores to measure pronunciation quality, e.g., correctness of manner and place of articulation. Moreover, the sub-segmental scores are then merged by neural networks to generate segmental scores.

2. MANDARIN PHONES & SPEECH ATTRIBUTES

In Mandarin, each Chinese character corresponds to one spoken syllable, consisting of an initial, usually a consonant, and a final, usually a vowel(s) or vowel(s) followed by a nasal. Speech attributes can be used to describe how consonant, vowel and nasal are produced using related articulators. Therefore, we can use such information to help detect initial and final mispronunciations. Table 1 lists the speech attribute categorization of Mandarin phones denoted in Pinyin. We adopt the same attribute-to-phone conversion rules as in [29][1].

Table 1. *Speech attributes and their associated phones in Pinyin*

Category	Attribute	Phone set
Place	Bilabial	B,P,M
	Labiodental	F
	Alveolar	D,L,N,T
	Dental	C,S,Z,I1
	Retroflex	ZH,CH,SH,R,ER,I2
	Palatal	J,Q,X,A,O,E,I,U,V
	Velar	G,H,K,NG
Manner	Stop	B,P,D,T,G,K
	Fricative	F,S,SH,R,X,H
	Affricative	Z,ZH,C,CH,J,Q
	Nasal	M,N,NG
	Lateral	L
	N/A	A,O,E,I,I1,I2,U,V,ER
Aspiration	Aspirated	P,T,K,C,CH,Q
	Unaspirated	B,D,G,Z,ZH,J
	N/A	F,H,L,M,N,R,S,SH,X,NG, A,O,E,I,I1,I2,U,V, ER
Voicing	Voiced	M,N,L,R,NG, A,O,E,I,I1,I2,U,V, ER
	Unvoiced	B,P,M,F,D,T,N,L,G,K,H,J,Q,X ZH,CH,SH,R,Z,C,S
Silence	Silence	SIL

3. OVERVIEW OF DETECTION FRAMEWORK

In this work, we adopt the automatic speech attribute transcription (ASAT) paradigm [27] to build the mispronunciation detection framework shown in Figure 1.

3.1. Attribute Feature Extraction

The feature extraction module consists of a bank of speech attribute classifiers. A context dependent DNN-based attribute classifier is separately built for each category described in Table 1. Expanded frames of input speech are fed into each detector, generating the current frame posteriors pertaining to each possible attribute within that category. Subsequently, a group of the frame attribute posteriors will be fed into the next module.

3.2. Sub-segmental Pronunciation Score Calculation

In this module, we adopt the goodness of pronunciation (GOP) calculation from [14]. Given unit p (e.g., context-independent attribute), we use Eq. (1) to calculate its log posteriors:

$$\log P(p|\boldsymbol{o}; t_s, t_e) = \frac{1}{t_e - t_s} \sum_{t=t_s}^{t_e} \log \sum_{s \epsilon p} P(s|\boldsymbol{o_t}), \qquad (1)$$

where $\boldsymbol{o_t}$ is the input feature at frame t; t_s and t_e are the start and end times of unit p, obtained by forced-alignment. $P(s|\boldsymbol{o_t})$ is the frame-level posterior; s is the context-dependent label (e.g., context-dependent manner attributes); $\{s \in p\}$ is the set of context-dependent units, whose central unit is p. Consequently, the GOP score for unit p is evaluated as

$$GOP(p) = \log \frac{P(p|\boldsymbol{o}; t_s, t_e)}{\max_{\{q \in Q\}} P(q|\boldsymbol{o}; t_s, t_e)}, \qquad (2)$$

where p is the canonical unit, q is the competing unit, and Q is the set of possible units within each category. A threshold is needed to

[1] The phoneme /I/ has 3 allophones: [I1] when followed by C, Z, S, [I2] when followed by ZH, CH, SH, R, and [I] when followed by all other initials.

verify whether the current unit is correctly pronounced. When computing unit is set to attribute, the log posterior of any given attribute can be calculated by using equation (1). We call these logarithmic posterior scores as sub-segmental scores.

3.3. Segmental Pronunciation Score Calculation

After being appended and expanded, sub-segmental scores are used to discriminate phone classes. The merger can be implemented with neural networks. The log-ratio between posterior of canonical phone p and that of the most competing phone q is adopted as the segmental score. To provide segmental level feedbacks (e.g. phone substitution), given one speech segment, the detection result in this study is phone p or q (depending on if the above ratio is larger or smaller than the threshold*)*. In the next experiment section, we enumerate a variety of possible thresholds to calculate the detection accuracy and display the precision-recall curves. As our preliminary study, this paper focuses on the mispronunciation of initials, because initial errors are more prone to cause miscommunication in Mandarin when compared to finals [30].

4. EXPERIMENTS

4.1. Speech Corpora

The native speech corpus is from the Chinese National Hi-Tech Project 863 for Mandarin LVCSR system development [31]. A total of 94,000 utterances spoken by 160 speakers (100 hours) were used for acoustic modeling.

The non-native speech corpus used is a subset of iCALL [25], containing 90,841 utterances spoken by 305 beginning learners of Mandarin. Each learner was asked to read Pinyin prompts, ranging from short phrases to sentences. All audio recordings are manually transcribed (surface pronunciation) by trained labelers, while the original Pinyin prompts were used as canonical pronunciation. By comparing the above surface and canonical transcriptions, we can get mispronunciation types at the phone level. Based on Table 1, after mapping each phone to its corresponding attributes, we can also identify mispronunciation types at the sub-segmental level.

The iCALL corpus is divided into 2 subsets: (i) training set (270 speakers), which was incorporated with the native training data to obtain the acoustic models; the training set only included short utterances, which had fewer mispronunciations compared to the long sentences [25]; and (ii) test set (30 speakers), which was used for evaluation. There is no speaker overlap between the training and test sets. Moreover, our L2 test set is made up of 5 different L1s, including English, French, Spanish, Italian and Russian. Such L1 diversity makes mispronunciation detection more challenging, because the error types made by different L2 learners are influenced by their L1s [7].

4.2. Phone-based Systems Setup

To better appreciate the empirical evidence gathered in the next sections and establish a sound assessment of the proposed framework, we implemented two phone-based systems, phone-based GOP system and phone-based classifier system inspired by [14], to compare with our proposed attribute-based classifier system. We should remark that the above two phone-based systems were trained and evaluated on our datasets and not those used in [14]. Although minor implementations differences might exist between our implementation and the MSRA systems in [14], the overall accuracy should not be heavily impacted. Therefore, we consider that the proposed one-to-one comparison fair.

Regarding phone-based GOP system, the training set from the native speech and non-native speech corpora are combined to train the CD-DNN-HMM acoustic models, which are used to calculate segmental pronunciation scores, where the unit p in Eqs. (1) and (2) is set to phone. The acoustic feature used to train the CD-DNN-HMM systems is 39-dim MFCC+Δ+ΔΔ. The DNN has six hidden layers, each with 2048 sigmoid units. The softmax function was employed at the output layer whose target classes are 2160 senone states [32] obtained from CD-GMM-HMM trained with a maximum likelihood criterion. The input to all the DNNs used in this paper is an augmented 11-frame vector, including 5 preceding, the current and 5 succeeding frames. A series of ASR experiments on the test set of the iCALL corpus is carried to assess the quality of this acoustic model. A free syllable loop grammar was used in the decoding process. The size of syllable counts is 420. Due to many confusable syllable pairs, the syllable recognition error rate is 34.71%, comparable to that achieved in [33].

After generating phone log-posteriors, augmented phone-level feature vectors were fed into the phone-based classifier system like our merger in the Figure 1. The difference between [14] and our implementation is that the output layer' labels change from two class labels (correct vs incorrect phone) to phone labels in order to provide a segmental-level feedback, e.g., phone substitution, in addition to detecting mispronunciations. After evaluating different number of hidden layers (1 or 2) and hidden nodes (512, 1024, and 2048) for each layer, we observed that 1 hidden non-linear layer with 1024 units gave the best performance on the training data.

4.3. Attribute-based Classifier System Setup

The input feature in Figure 1 is a 39-dim MFCC+Δ+ΔΔ vector. After forced-alignment, context dependent (CD) attribute labels are separately used to train corresponding DNNs containing 6 hidden layers each having 2048 sigmoid units. The softmax function was employed at the output layer. Each attribute classifier generates CD acoustic-phonetic feature (attribute) posteriors (e.g., stop-vowel+nasal). These probabilities are fed into the GOP calculator to compute the sub-segmental scores, which are subsequently used for sub-segmental mispronunciation detection. Finally sub-segmental scores are merged into segmental scores by neural networks to detect mispronunciation at the segmental level, namely at a phone level. After evaluating the different number of hidden layers (1 or 2) and hidden nodes (512, 1024, and 2048) for each layer, we observed that 1 hidden non-linear layer with 1024 units gave the best performance on the training data.

4.4. Evaluation Metrics

In this study, four metrics, namely precision, recall, detection accuracy (DEA) and diagnostic accuracy (DA), are used to evaluate the performance of each mispronunciation system:

$$Precision = \frac{N_M}{N_D} * 100\% \quad (3)$$

$$Reacll = \frac{N_M}{N_H} * 100\% \quad (4)$$

$$DEA = \frac{N_M + N_C}{N} * 100\% \quad (5)$$

$$DA = \frac{N_F}{N_M} * 100\% \quad (6)$$

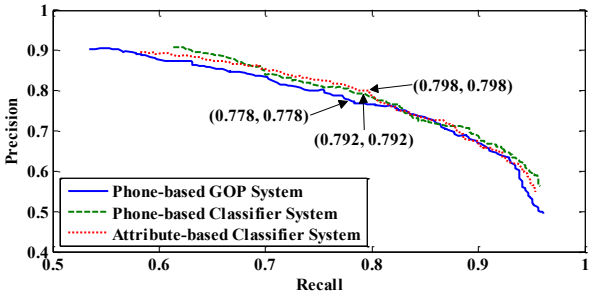

Figure 2: *Comparing segmental detection performance*

where N_M is the number of true mispronunciations detected, N_D is the total number of detected mispronunciations, N_H is the total number of mispronunciations labeled by a human expert, N_C is the number of true correct pronunciation detected by the system; N_F is the number of truly detected mispronunciation, where feedback is correct, and N is the number of phone or attribute in the test set.

4.5. Experimental Results

Table 2 shows 1-EER (equal error rate), and the sub-segmental mispronunciation detection performance, DEA and DA, at the EER operating point when precision equals recall. Table 3 compares three systems: our attribute-based classifier system, phone-based GOP and classifier systems, at the segmental level. Figure 2 shows precision recall curves and the 1-EER points.

Table 2. *Detection and diagnostic accuracy at the EER operating point where precision is set the same as recall (sub-segmental)*

	1-**EER**	DEA	DA
VOICING	77.68%	99.4%	100%
ASPIRATION	77.35%	95.7%	96.3%
MANNER	73.92%	96.7%	97.0%
PLACE	63.92%	95.5%	93.9%

Table 3. *Detection and diagnostic accuracy where precision is set to be the same as recall (segmental level)*

	1-**EER**	DEA	DA
PHONE-BASED GOP SYSTEM	77.80%	90.84%	86.08%
PHONE-BASED CLASSIFIER SYSTEM	79.20%	91.32%	86.99%
ATTRIBUTE-BASED CLASSIFIER SYSTEM	79.75%	91.57%	87.01%

4.6. Discussions

In this work, we examine mispronunciations in five categories. Four of which are at the sub-segmental level, and the remaining one at the segmental level. For each category, a single threshold is used to determine the correctness of non-native pronunciations. We discuss Figure 2, Table 2 and Table 3 below.

Regarding the segmental level, our proposed attribute–based classifier system outperforms the MSRA-style phone-based GOP system with a relative EER reduction of 8.78% (see Figure 2). The corresponding detection and diagnostic errors were relatively reduced by 7.97% and 6.68%, respectively. Furthermore, compared with standard phone-based classifier systems, the proposed attribute-based approach attains competitive performance as shown in Figure 2, while providing useful sub-segmental level feedback, which will be illustrated in the following.

Unlike phone modeling units, our proposed speech attributes can be used to detect mispronunciation at the sub-segmental level. From Table 2, we can leverage upon the high DEA and DA rates to deploy attributes for mispronunciation detection and feedback. Figure 3 is an example using the place of articulation units to detect mispronunciations. The upper panel plots the spectrogram of one syllable and its corresponding canonical phone sequence. Compared with its surface phone sequence, we find that the unaspirated retroflex affricate phone /zh/ is mispronounced to its dental counterpart /z/. Our speech attribute detection results shown in the lower panel find that the retroflex posterior is much lower than the dental posterior. However the place of articulation of the canonical phone /zh/ should be retroflex. Reflecting on the above observations, our sub-segmental feedback could be formulated as "*Try to move your tongue tip backwards so that the edges of your tongue are touching your hard palate*". This feedback at the sub-segmental level is specific and can enrich the traditional diagnostic information, e.g., phone substitution.

Different sub-segmental categories have different detection performances, e.g., voicing achieves the best EER, DEA and DA, as shown in Table 2, while place of articulation is less accurate as indicated. The performance discrepancy is partly due to the fact that there are more attribute classes for place of articulation. In addition, Mandarin affricates and fricatives differ subtlety in terms of acoustics in the coronal region (e.g., alveolar, retroflex, palatal).

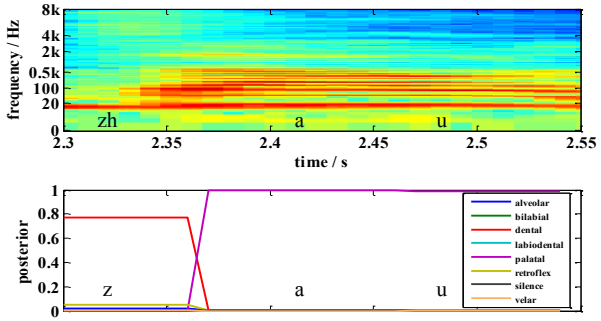

Figure 3: *Sub-segmental mispronunciation detection using speech attributes (place of articulation). The canonical and surface transcriptions are shown in the upper and lower panel respectively.*

5. CONCLUSION AND FUTURE WORK

In this paper, speech attribute modeling is proposed to provide comprehensive diagnostic feedback at the segmental and sub-segmental level for non-native mispronunciations. Compared with the phone-based baselines, our proposed attribute-based system achieves better detection results than GOP-based approaches and comparable detection results with classification-based approaches at the segmental phone level with an advantage of speech attribute portability across different languages [34]. For future work, the detection errors produced by phone-based and attribute-based systems will be analyzed. Systems combination will also be investigated. Furthermore, a recent DNN extension to GOP [35] and similar DNN-based utterance verification [7-11] extensions can also be utilized to improve frame-level and segment-level pronunciation verification.

6. REFERENCES

[1] D. Graddol, "Why global English may mean the end of English as a Foreign Language," *ULIS*, 2008.

[2] "40 million people worldwide study Chinese," http://english.people.com.cn/90001/90782/90872/7112508.html.

[3] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA, *MIT Press*, 2000.

[4] G. Fant, *Speech Sounds and Features*. Cambridge, MA, *MIT Press*, 1973.

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

[6] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[7] R. A. Sukkar and C.-H. Lee, "Vocabulary Independent Discriminative Utterance Verification for Non-Keyword Rejection in Subword Based Speech Recognition," *IEEE Trans. on Speech and Audio Proc*., Vol. 4, No. 6, pp. 420-429, Nov. 1996.

[8] M. Rahim, C.-H. Lee, and B.-H. Juang, "Discriminative Utterance Verification for Connected Digit Recognition," *IEEE Trans. on Speech and Audio Proc*., Vol. 5, No. 3, pp. 266-277, May 1997.

[9] Sukkar, A. R. Setlur, C.-H. Lee and J. Jacob, "Verifying and Correcting String Hypotheses Using Discriminative Utterance Verification," *Speech Communication*, Vol. 22, pp. 333-342, 1997.

[10] T. Kawahara, C.-H. Lee and B.-H. Juang, "Key-Phrase Detection and Verification for Flexible Speech Understanding," *IEEE Trans. on Speech and Audio Proc*., Vol. 6, No. 6, pp. 558-568, Nov. 1998.

[11] M.-W. Koo, C.-H. Lee and B.-H. Juang, "Speech Recognition and Utterance Verification Based on a Generalized Confidence Score," *IEEE Trans. on Speech and Audio Proc*., Vol. 9, No. 8, pp. 821-832, Nov. 2001.

[12] J. Zheng, C. Huang, M. Chu, F. K. Soong, and W. Ye, "Generalized Segment Posterior Probability for Automatic Mandarin Pronunciation Evaluation," in *Proc. ICASSP*, 2007.

[13] S. Wei, G. Hu, Y. Hu, R.H. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models," *Speech Communications*, vol. 51, no. 10, pp. 896–905, 2009.

[14] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved Mispronunciation Detection with Deep Neural Network Trained Acoustic Models and Transfer Learning based Logistic Regression Classifiers," *Speech Communication*, 67, pp. 154-166, 2015.

[15] H. Meng, Y. Lo, L. Wang, and W. Yiu, "Deriving salient learners' mispronunciations from cross-language phonological comparisons," in *Proc. ASRU*, 2007.

[16] X. Qian, H. Meng, and F. K. Soong, "Capturing L2 Segmental Mispronunciations with Joint-sequence Models in Computer-Aided Pronunciation Training (CAPT)," in *Proc. ISCSLP*, 2010.

[17] 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.

[18] C. D. Chu, and N. F. Chen, "Stop-like modification of dental fricatives in Indian English: A preliminary study to perceptual experiments," *J. Acoust. Soc*. Am. 125, 2778, 2009.

[19] O. Bälter, O. Engwall, A. Öter, and H. Sidenbladh-Kjellström, "Wizard-of-Oz test of ARTUR: a computer-based speech training system with articulation correction," in *Proc. of the Seventh International ACM SIGACCESS Conference on Computers and Accessibility*, pages 36–43.

[20] J. Tepperman and S. Narayanan, "Using articulatory representations to detect segmental errors in nonnative pronunciation," in *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):8-22, Jan. 2008.

[21] O. Engwall, "Pronunciation analysis by acoustic-to-articulatory feature inversion," in *Proc. of the International Symposium on Automatic Detection of Errors in Pronunciation Training*, 2012.

[22] O. Engwall, "Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher," *Computer Assisted Language Learning*, 25:37–64, 2012.

[23] J. Tepperman and S. Narayanan, "Hidden-articulator Markov models for pronunciation evaluation," in *Proc. ASRU*, 2005.

[24] X. Xie, and W. Abdulla, "Computer Aided Pronunciation Learning Systems," *(Computer Systems Engineering)--University of Auckland*, 2010.

[25] 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.

[26] D. Yu, S. Siniscalchi, L. Deng, and C.-H. Lee, "Boosting attribute and phone estimation accuracies with deep neural networks for detection based speech recognition," in *Proc. ICASSP*, 2012.

[27] C.-H. Lee and S. M. Siniscalchi, "An information-extraction approach to speech processing: Analysis, detection, verification, and recognition," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1089– 1115, 2013.

[28] Li K, Qian X, Kang S, et al, "Integrating Acoustic and State-Transition Models for Free Phone Recognition in L2 English Speech Using Multi-Distribution Deep Neural Networks," in *Proc. SLaTE*, 2015.

[29] C. Zhang, Y. Liu, and C.-H. Lee, "Detection-based accented speech recognition using articulatory features," in *Proc. ASRU*, 2011.

[30] J.-S. Zhang, W. Li, et al, "A Study On Functional Loads of Phonetic Contrasts Under Context Based On Mutual Information of Chinese Text And Phonemes," in *Proc. ISCSLP*, 2010.

[31] S. Gao, et al, "Update of Progress of Sinohear: Advanced Mandarin LVCSR System At NLPR," in *Proc. ICSLP*, 2000.

[32] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition," *IEEE Trans. on Speech and Audio Proc*, vol. 20, no. 1, pp. 33-42, 2012.

[33] R. Tong, N. F. Chen, B. Ma, H. Li, "Goodness of Tone (GOT) for Non-native Mandarin Tone Recognition," in *Proc. Interspeech*, 2015.

[34] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Universal attribute characterization of spoken languages for automatic spoken language recognition," *Computer Speech and Language*, vol. 27, no. 1, pp. 209–227, 2013.

[35] 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.