Mandarin Mispronunciation Detection and Diagnosis Feedback Using Articulatory Attributes Based Multi-task Learning

Xuan-Bo Chen*, Yueh-Ting Lee*, Hung-Shin Lee†, Jyh-Shing Roger Jang*, Hsin-Min Wang†
*Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
†Institute of Information Science, Academia Sinica, Taiwan
bruce.chen@mirlab.org, ayueh.lee@mirlab.org

***Abstract*—This paper presents our research on computer assisted pronunciation training (CAPT). We focus on mispronunciation detection and articulation feedback. We propose taking into account the speech attributes, namely place and manner of articulation, in the assessment models to improve mispronunciation detection and return precise articulation feedback to learners. We train a discriminative articulatory model based on time-delay neural networks (TDNNs) with the multi-task learning strategy to give the articulatory score and a TDNN-based acoustic model to give the phonetic score. In testing, the system detects mispronunciations and returns precise articulation feedback based on both the phonetic and articulatory scores. The results of experiments conducted on the MATBN Mandarin Chinese broadcast news corpus show that the proposed models outperform the Gaussian mixture model (GMM)-based and deep neural network (DNN)-based baselines in terms of equal error rate (EER) and diagnostic accuracy (DA). Furthermore, our mispronunciation detection system should work in any language, although the current system focuses on Mandarin.**

***Index Terms*—computer assisted pronunciation training, mispronunciation detection, articulatory features, multi-task learning, discriminative training, time-delay neural networks.**

I. Introduction

Computer assisted pronunciation training (CAPT) systems provide opportunities for learners to practice pronunciation in a stress-free environment. In the past few decades, CAPT systems based on statistical modeling techniques have made considerable progress [1–4]. For effective learning, CAPT systems should provide learners with pronunciation assessments and personalized correction feedback. In general, there are two main approaches to pronunciation assessment. One is to give learners pronunciation scores, from the phone level to the syllable level [5, 6], and the other is to detect individual errors, such as specific phone substitution errors [7, 8]. Based on the scores at the phone or syllable level, learners can know how well their pronunciation is and where the wrong pronunciation might be, but they cannot know the types of errors and how to correct them. Concerning the detection of phone substitution errors, researchers usually target a few specific problematic phones. For example, a typical CAPT system will give the substitution feedback on the "r-l substitution error", which occurs very commonly in English when a learner pronounces the word "rate" as "late". Rather than providing such phone substitution feedback, giving feedback directly related to corrective articulation from articulatory model is more useful to learners. Articulatory models can be categorized into geometrical [9–12] and biomechanical [13, 14] types. In this paper, we focus on the geometrical model. In a geometrical model, the vocal tract is represented by its initial geometry, and a set of parameters estimated from the electromagnetic articulography (EMA) data directly deforms this geometry. Joint mispronunciation detection and articulation suggestions has been proved helpful in many areas, such as speech therapy [15], speech comprehension improvement [16] and pronunciation perceptual training [17].

In this study, we propose a mispronunciation detection system that takes into account the phonetic and articulatory scores at the phone and syllable levels and returns precise articulation feedback to learners. We perform experiments on the MATBN Mandarin Chinese broadcast news corpus [18], where all the speech utterances were manually annotated with transcripts, speakers, and syllable mispronunciations. We use a subset of the MATBN corpus spoken by speakers who have no mispronunciations in their speech to train a time-delay neural network (TDNN)-based acoustic model by Kaldi[1] for phonetic forced alignment and phonetic score evaluation [19]. We also use the same subset to train a TDNN-based speech attribute (i.e., manner and place attributions of articulation) classifier with the multi-task learning strategy for articulatory score evaluation. In testing, the phonetic scores are calculated by the acoustic model while the articulatory scores are calculated by the articulatory model (i.e., the attribute classifier). The system detects mispronunciations and returns articulation feedback based on both the phonetic and articulatory scores. Another subset of the MATBN corpus that contains mispronunciations is used as the testing data, mostly the interviewees' speech.

In summary, the highlights of this paper in comparison with other relevant reported works are threefold.

1) TDNNs are employed in our CAPT system instead of fully-connected multilayer perceptrons to efficiently learn the temporal dynamics of signals from short-term feature representations [20, 21].

[1]http://kaldi-asr.org

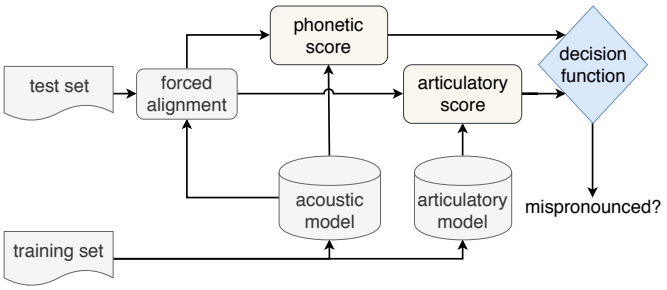

Fig. 1. Overview of the proposed mispronunciation detection framework.

2) The i-Vector, as one of input components of TDNNs, is used to normalize the variability of speakers [22].
3) The output units in each articulatory model are simplified to mono-attributes rather than tri-attributes used in [23]. The attributes in each attribute category are mutually exclusive.

The remainder of this paper is organized as follows. Section II introduces the speech attributes for Mandarin phones. Section III describes our mispronunciation detection framework and methods of scoring the context-dependent phones. Section IV presents our training strategy for the attribute classifier. Section V gives an overview of the corpus used in this paper. Section VI reports the experimental results. Finally, conclusions are drawn in Section VII.

II. Speech Attributes for Mandarin Phones

In Mandarin, each Chinese character corresponds to one spoken syllable, consisting of an initial, usually a consonant, and a final, usually vowel(s) or vowel(s) followed by a nasal. Speech attributes can be used to describe how phones are produced using related articulators and the airflow from the lungs, and thus can be used to detect changes in pronunciation caused by either regional accent or substitution errors. In this study, we detect mispronunciations based on the phone-level posteriors and speech attributes. We adopt the attribute-to-phone conversion rules in [24][2]. Table I lists the different categories of speech attributes and their associated Mandarin grapheme-phoneme denoted in Hanyu Pinyin.

According to [23], place and manner of articulation are used to describe the attributes of consonant sounds, while vowels are described with three-dimensional features: horizontal dimension (tongue backness), vertical dimension (tongue height), and lip shape (roundedness). Based on Table I, the articulatory attribute transcription of a speech utterance can be directly derived from its phone transcription. In this study, we use four kinds of articulatory transcriptions: manner, place + backness, place + height, and place + roundedness. Table II gives an example of attribute labels derived from the phone labels. By forced alignment with a pre-trained phone-based acoustic model, each frame of a speech utterance can be labeled with the phone and the articulatory attributes represented by four ground truth one-hot vectors.

[2]The phoneme `i` has 3 allophones: `ii` when followed by `c`, `z`, `s`, `iii` when followed by `zh`, `ch`, `sh`, `r`, and `i` when followed by all other initials.

TABLE I
Five categories of speech attributes and their associated Mandarin grapheme-phonemes in Hanyu Pinyin.

category	attribute	grapheme-phoneme
place	bilabial	`b p m`
	labiodental	`f`
	alveolar	`d t l n`
	dental	`z c s`
	retroflex	`zh ch sh r`
	palatal	`j q x`
	velar	`g k h`
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`
	lateral	`l`
	n/a	all vowels
backness	back	`o er u`
	central	`a err iii`
	front	`e i v ii nn ng`
	n/a	all consonants
height	high	`i ii iii u v`
	low	`a ng`
	middle high	`o er nn`
	middle low	`e err`
	n/a	all consonants
roundedness	rounded	`o u v ng`
	unrounded	`a er e err i ii iii nn`
	n/a	all consonants

TABLE II
The phoneme sequence and its corresponding manner, place + backness, place + height, and place + roundedness sequences, taking the Chinese phrase "我們" (we) for example.

phrase	**我們 (we)**				
phoneme	`u`	`o`	`m`	`er`	`nn`
manner	vowel	vowel	nasal	vowel	vowel
place + backness	back	back	bilabial	back	front
place + height	high	middle high	bilabial	middle high	middle high
place + roundedness	rounded	rounded	bilabial	unrounded	unrounded

III. Overview of the Proposed Mispronunciation Detection Framework

Our mispronunciation detection framework is shown in Fig 1. We first train an acoustic model and an articulatory model from a well-pronounced speech corpus. Given a test speech utterance, we first perform forced alignment to divide it into a phonetic segment sequence. Then, we compute the phonetic score and the articulatory score for each phonetic segment. Finally, the decision module will judge whether a phonetic segment or a syllable segment is mispronounced. We will describe the individual modules in detail in this section.

A. The phonetic score calculated by the acoustic model

We can use the goodness of pronunciation (GOP) [6] as the phonetic score. The GOP for a given phonetic segment O^p labeled as phone p is defined as the conditional probability $P(p|O^p)$, i.e., the posterior probability of p given the phonetic segment O^p. Therefore, the phonetic score ζ_p of a phonetic segment O^p for phone p is calculated as follows:

$$\zeta_p = \left| log\left(\frac{P(O^p|p)}{\max_{q\in Q} P(O^q|q)}\right)\right| /NF(p), \tag{1}$$

where q is the competing phone; Q is the complete phone set, and $NF(p)$ is the number of frames in the phonetic segment O^p. GOP is in the range of (-∞ ,0]. The best GOP is zero when $p = q$, the bigger the better. A threshold is demanded to verify whether the phonetic segment is correctly pronounced.

We can also calculate the phonetic score with the rank ratio (RR) [25], which is computes as follows:

$$\zeta_p = \frac{R_p - 1}{N}, \tag{2}$$

where R_p denotes the rank of the log-likelihood of the acoustic model for phone p and N is the number of phones in the complete phone set. RR is in the range of [0, 1). The best RR is zero when $R_p = 1$, the smaller the better.

1) Normalization of the phonetic score: As introduced above, the RR phonetic score is in the range of [0, 1) while the GOP phonetic score is in the range of (-∞ ,0]. For both types of scores, a score close to zero indicates a good pronunciation. Therefore, we can normalize these two scores to the same range between 0 and 100, and the larger the value, the better the pronunciation. The normalization function is as follows:

$$\widetilde{\zeta_p} = \frac{100}{1 + (\frac{|\zeta_p|}{a})^b}, \tag{3}$$

where $\widetilde{\zeta_p}$ denotes the normalized phonetic score for phone p; ζ_p can be the phonetic score obtained from either Eq. (1) or Eq. (2); and a and b are constants. In this study, we set a to 0.2 and b to 2. Note that, for the sake of simplicity, we will use ζ_p to represent the normalized phonetic score hereafter.

2) The syllable-level phonetic score: To calculate the syllable-level phonetic score, we can simply compute the average of the phonetic scores of the component phones in the syllable. Alternatively, we can calculate the syllable-level score in line with the phone lengths by the result of forced alignment. That is, we regard the phonetic segment length (i.e., the number of frames) as a weight of the corresponding phonetic score in calculating the syllable-level score as follows:

$$\zeta_s = \frac{1}{L}\sum_{i}^{n}(L_i * \widetilde{\zeta_i}), \tag{4}$$

where ζ_s denotes the phonetic score of syllable s, i denotes the i^{th} phone in syllable s, L_i denotes the segment length of the i^{th} phone, $\widetilde{\zeta_i}$ denotes the normalized phonetic score of the i^{th} phone, n is the number phones in syllable s, and $L = \sum_i^n L_i$.

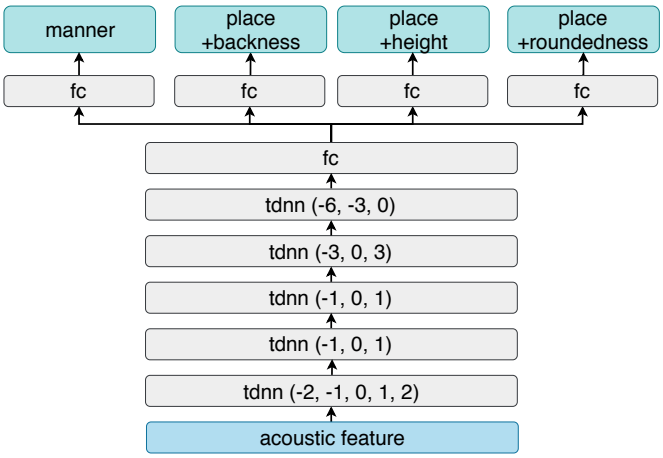

Fig. 2. The architecture of the TDNN-based attribute classifier, where $tdnn$ and fc denote the time-delay layer and fully-connected layer, respectively. The notation followed by $tdnn$ represents the specific layer-wise context.

TABLE III
THE GROUND TRUTH MANNER ATTRIBUTE VECTOR AND THE PREDICTED MANNER ATTRIBUTE VECTOR FOR A SPEECH FRAME OF "M" PHONE. "PREDICTED" MEANS THE OUTPUT OF THE MANNER LAYER IN FIG 2.

attributes	**stop**	**fricative**	...	**nasal**
ground truth	0	0	...	1
predicted	0.01	0.03	...	0.95

B. The articulatory score calculated by the articulatory model

As shown in Fig. 2, our TDNN-based attribute classifier will output the predicted manner, place + backness, place + height, and place + roundedness vectors for each frame of input speech. Since each frame corresponds to a specific phone according to the forced alignment by the acoustic model, the ground truth manner, place + backness, place + height, and place + roundedness vectors can be easily derived according to Table II. In Table III, we take the phoneme `m` for example to explain the ground truth vectors and the predicted vectors. The first row in Table III represents the ground truth manner vector of an `m` frame, which is a one-hot vector with value "1" for the nasal attribute. The second row in Table III is the predicted manner vector. We can compute the inner product of these two vectors as the articulatory score as follows:

$$\eta_p = \frac{1}{L}\sum_{i}^{L}\frac{[\widehat{att.}_i]\cdot[att._i] * 100}{4}, \tag{5}$$

where η_p denotes the articulatory score of phone p; i denotes the i^{th} frame of the phone p segment; L denotes the number of frames in the phone p segment; $\widehat{att.}_i$ and $att._i$ denote the ground truth attribute vector and the predicted attributed vector of the i^{th} frame, respectively. Note that $\widehat{att.}_i$ and $att._i$ are the concatenated vector of the 4 ground truth attribute vectors and the concatenated vector of the 4 predicted attribute vectors, respectively. Therefore, in Eq. (5), the 4 at the denominator is to normalize the inner product between 0 and 1 while the 100 at the numerator is to scale the articulatory score to the same interval as the phonetic score for consistency.

C. Combination of the phonetic and articulatory score

For phone p, the overall assessment score can be simply calculated as the weighting average of the phonetic score ζ_p and the articulatory score η_p as follows:

$$\lambda_p = w * \zeta_p + (1 - w) * \eta_p, \tag{6}$$

where w is an adjustable weight.

IV. Multi-task learning

As discussed above, each frame of a speech utterance can be labeled with phone and four kinds of ground truth one-hot attribute vectors. We can train a TDNN-based attribute classifier for each attribute category separately [26]. In this study, we take advantage of multi-task learning and propose attribute modeling that considers all the interaction effects in a unified objective function in Eq. (7). Fig. 2 shows the TDNN-based attribute classifier built for the four kinds of attribute vectors presented in Sec. II.

A. The advantage of multi-task learning

The purpose of using multi-task learning [27, 28] (MTL) is to train the model with several different but related tasks using shared representations. The effectiveness of MTL depends on the relationship between each task and the shared learning structure across tasks. In our case, the four output layers correspond to four categories of speech attributes, respectively, so they are highly connected to each other. Not only the different articulatory features can be modeled at the same time, but also each feature can be prevented from over-fitting.

B. The neural network architecture

Since the attributes in each of the four attribute vectors are exclusive, and the four attribute vectors are related, we train our TDNN-based attribute classifier with multi-task learning. Specifically, each speech frame in the input layer is represented by a vector consisting of 100 i-Vector features, 40 high resolution MFCCs, and 3 pitch features. The first five hidden layers in Fig. 2 are time-delay layers [21]. $(-3, 0, 3)$ means that we splice frames $\{t-3\}$, $\{t\}$, and $\{t+3\}$ together at the current hidden layer and t denotes the current time. The last two hidden layers, labeled as fc in Fig. 2, are fully-connected layers with 650 hidden nodes for the shared fc layer and 250 hidden nodes for the fc layer of each sub-task. The objective function of multi-task learning is the summation of the cross-entropy losses of the four sub-tasks:

$$\begin{aligned} Loss = & CE_{\text{manner}} + CE_{\text{place+backness}} + \\ & CE_{\text{place+height}} + CE_{\text{place+roundedness}}, \end{aligned} \tag{7}$$

where $Loss$ and CE denote the total loss of the TDNN-based attribute classifier and the cross-entropy loss of the output layer of an individual sub-task.

TABLE IV
STATISTICS OF DIFFERENT TYPES OF MISPRONUNCIATIONS IN MATBN. IN THIS STUDY, WE FOCUS ON PHONE SUBSTITUTION ERRORS, I.E., THE INITIAL PHONE ERROR, MIDDLE PHONE ERROR, AND FINAL PHONE ERROR. THE OTHER ERROR DENOTES PHONE INSERTIONS AND DELETIONS. THE TONE ERROR MEANS IT IS PRONOUNCED WITH THE INCORRECT TONE. THE UNDERLINED PHONE REPRESENTS THE LOCATION OF MISPRONUNCIATION.

error type	proportion (%)	top pattern	proportion (%)
initial phone error	69.09	shiii → sii	13.09
middle phone error	2.33	xve → xie	2.05
final phone error	3.37	qv → qi	0.48
other error	18.40	fa → hua	10.56
tone error	6.81	fu2 → fu3	1.16

V. Speech corpus

We use the MATBN corpus, which is a Mandarin Chinese broadcast news corpus collected in Taiwan [18], in the experiments. There are three types of speakers: anchor reporters, field reporters, and interviewees. The speech of anchor reporters always exhibits a high standard of fluency, good pronunciation and good acoustic quality. The speech of field reporters also exhibits a high standard of fluency and good pronunciation but sometimes the acoustic quality is low. The speech of interviewees is often of very low quality and intelligibility with background speech and noises of various types. All the speech utterances in the corpus are manually annotated with transcripts, speakers, and syllable mispronunciations.

A. Mispronunciations in the speech corpus

The statistics of the mispronunciations in the MATBN corpus are summarized in Table IV. There are two major categories of mispronunciations. The first corresponds to the error in tone (e.g., `fu2` pronounced as `fu3`), and the second corresponds to the phone errors (e.g., `sha1` pronounced as `sa1`), including substitution, insertion, and deletion. We can see that the mispronunciation of consonants (cf. the initial phone error in Table IV) accounts for the largest proportion of mispronunciations. In this study, we only consider the phone substitution errors, i.e., the initial phone error, the middle phone error, and the final phone error, according to the location of the error. We leave other errors to future research.

B. Training and testing sets

Among the three types of speakers, studio anchors and field reporters are well-trained native speakers. Therefore, there are few mispronunciations in their speech. In contrast, there are more pronunciation errors in the interviewees' speech.

We thought that the acoustic model for pronunciation assessment should be trained by a standard speech corpus without mispronunciations, such as the speech of well-trained anchor and field reporters. The speech with mispronunciations could then be used as the testing data. Consequently, we divided the MATBN corpus into two subsets, namely MATBN-STD and MATBN-MIS. As shown in Table V, MATBN-STD

TABLE V
STATISTICS OF THE MATBN-STD AND MATBN-MIS SUBSETS. MATBN-STD CONTAINS ONLY THE CORRECTLY PRONOUNCED SPEECH FROM THE ANCHORS AND FIELD REPORTERS, AND MATBN-MIS CONTAINS THE SPEECH FROM THE ANCHORS, FIELD REPORTERS, AND INTERVIEWEES WITH MISPRONUNCIATIONS.

role	MATBN-STD		MATBN-MIS	
	#utt.	dur. (hours)	#utt.	dur. (hours)
anchor	5,599	22.4	693	3.8
reporter	10,211	41.3	3,153	16.8
interviewee	0	0	5,451	22.9
total	15,810	63.7	9,297	43.5

consists of 15,810 utterances (63.7 hours), and is used for training the acoustic model. MATBN-MIS consists of 9,297 utterances (43.5 hours), and is used for testing. Note that the speakers in training and testing subsets are not overlapped.

VI. EXPERIMENTS

We use the phone-based symbols to represent the phone set. We consider a syllable that is incorrectly pronounced as a positive example, and treat its left and right syllables as negative examples. If the left or right syllable around the mispronounced syllable is also mispronounced, we go forward or backward to find the nearest correctly pronounced syllable as the negative example. Therefore, the ratio of positive and negative examples is about 1:2 in the testing data.

A. Systems compared in the paper

We consider two types of transcripts, the base syllable sequence and the tonal syllable sequence. For example, the syllables `ba1`, `ba2`, `ba3`, `ba4`, and `ba5` are treated as the same base syllable `ba`. We add the tone label to the vowels and the last phone in a syllable.

We compare two acoustic models in the paper:

1) GMM: the baseline using the conventional GMM-HMM model as the acoustic model.
2) TDNN: the proposed method using the TDNN model as the acoustic model.

As introduced in Sec. III, two phonetic scores (i.e., RR, GOP) can be calculated by the acoustic model. In addition, experiments are conducted on without-tone and with-tone transcripts separately to examine their respective performance.

In our preliminary analysis, we found that the retroflex phones were difficult to recognize. For example, the phone `sh` was often pronounced as `s`, but the phonetic score of `sh` was still high. Furthermore, in the mispronunciation annotations, the retroflex errors are in the majority as shown in Table. IV (cf. the top pattern in the initial phone error category). Therefore, we came up with an idea to combine the acoustic and articulatory scores on the retroflex phones, as described in Algorithm 1.

B. Evaluation metrics

In this paper, the equal error rate (EER) and diagnostic accuracy (DA) are used for performance evaluation. In the confusion matrix in Table VI, the positive label stands for the mispronunciation. When a phone is predicted as positive by the system, if the phone with the highest posterior probability is the same as the mispronunciation annotation or all the four articulatory vectors exactly match the mispronunciation annotation, it is counted as a full hit; otherwise, we refer it as a near hit.

The diagnostic accuracy is defined as

$$DA(\theta_{EER}) = \frac{\text{\# full hit}}{\text{\# true positive}}, \tag{8}$$

where θ_{EER} denotes the threshold in the condition of EER. Note that if two systems have the same EER, the system with a higher DA score means it has a better diagnostic ability.

Algorithm 1: Combing phonetic and articulatory scores

Data: ζ_p, η_p
Result: λ_p
ζ_p *is obtained from Eq.* (1) *or Eq.* (2);
η_p *is obtained from Eq.* (5);
$w = 0.2$ (*heuristically selected*);
while *not at the end frame in phone p* **do**
  read the current frame;
  if *phone p has the retroflex attribute* **then**
    $\lambda_p = w * \zeta_p + (1 - w) * \eta_p$;
  else
    $\lambda_p = \zeta_p$;
  end
end

TABLE VI
THE CONFUSION MATRIX. WE DEFINE THE POSITIVE LABEL AS THE MISPRONUNCIATION AND THE NEGATIVE LABEL AS THE CORRECT PRONUNCIATION, AND TRUE POSITIVE = FULL HIT + NEAR HIT.

mispronunciation annotation	system prediction	
	positive	negative
positive	full hit / near hit	false negative
negative	false positive	true negative

TABLE VII
PERFORMANCE IN TERMS OF EER AND DA FOR DIFFERENT SYSTEMS. +PM DENOTES THE SYSTEM THAT COMBINES THE PHONETIC AND ARTICULATORY SCORE DERIVED BY ALGORITHM 1.

method	without tone		with tone	
	EER (%)	DA (%)	EER (%)	DA (%)
RR (GMM)	40.65	14.38	45.97	11.46
RR (GMM) + PM	33.70	32.56	36.98	34.94
RR (TDNN)	39.56	16.76	42.71	13.46
RR (TDNN) + PM	**33.27**	**43.52**	**35.81**	**42.96**
GOP (GMM)	41.31	17.49	45.80	17.29
GOP (GMM) + PM	38.56	36.04	42.60	37.81
GOP (TDNN)	39.85	22.43	41.78	19.97
GOP (TDNN) + PM	**33.55**	**46.79**	**36.86**	**44.65**

C. Results

Table VII compares the performance of sixteen experimental setups in terms of EER and DA. Several observations can be

drawn from the table. First, combining the articulatory score with the phonetic score is helpful for all the systems, regardless of which model is used to calculate the phonetic score and which type of the phonetic score is used. Second, the TDNN-based acoustic model outperforms the GMM-based acoustic model in terms of both EER and DA in all cases. Third, RR seems to be more stable than GOP, and the GOP score given by GMM is not so reliable.

We found the common cases in false negative are retroflex phones. By considering the articulatory factor in scoring, more retroflex phone errors like `zh` $\rightarrow$ `z`, `ch` $\rightarrow$ `c`, and `sh` $\rightarrow$ `s` were detected. Moreover, we could enhance the traditional diagnosis feedback (e.g., phone-substitution) through articulatory information.

VII. Conclusions

In this paper, we have proposed using the articulatory factor in speech to furnish encyclopedic diagnosis feedback and demonstrate that it can effectively detect mispronunciations on the retroflex phones. Two different phonetic scoring criteria, namely RR and GOP, have similar performance in baseline experiments, but after using the TDNN model and combining the articulatory score, the RR-based criterion performs better than the GOP-based one. In our future work, we will collect and annotate more mispronunciations and use two acoustic models to deal with bad alignments caused by serious mispronunciations, one for alignment and one for scoring. In order to provide more comprehensive diagnosis feedback to learners, more types of mispronunciations should be considered. Therefore, we need to investigate the pronunciation errors common in Mandarin in more detail through linguistic theories.

References

[1] W. Menzel, D. Herron, P. Bonaventura, and R. Morton, "Automatic detection and correction of non-native English pronunciation," in *Proc. InSTIL*, 2000.

[2] S. Seneff, C. Wang, and J. Zhang, "Spoken Conversational Interaction for Language Learning," in *Proc. InSTIL/ICALL*, 2004.

[3] H. Strik, J. Colpaert, J. van Doremalen, and C. Cucchiarini, "The DISCO ASR-based CALL System: Practicing L2 Oral Skills and Beyond," in *Proc. LREC*, 2012.

[4] X. Qian, H. Meng, and F. Soong, "A Two-Pass Framework of Mispronunciation Detection and Diagnosis for Computer-Aided Pronunciation Training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 6, pp. 1020–1028, 2016.

[5] H. Franco, L. Neumeyer, Yoon Kim, and O. Ronen, "Automatic Pronunciation Scoring for Language Instruction," in *Proc. ICASSP*, 1997.

[6] S. M. Witt and S. J. Young, "Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning," *Speech Communication*, vol. 30, no. 2, pp. 95–108, 2000.

[7] A. Lee and J. Glass, "Mispronunciation Detection without Nonnative Training Data," in *Proc. Interspeech*, 2015.

[8] S. Joshi, N. Deo, and P. Rao, "Vowel Mispronunciation Detection Using DNN Acoustic Models with Cross-lingual Training," in *Proc. Interspeech*, 2015.

[9] S. Maeda, "Compensatory Articulation During Speech: Evidence from the Analysis and Synthesis of Vocal-tract Shapes Using an Articulatory Model," in *Speech Production and Speech Modelling*. Springer, 1990, pp. 131–149.

[10] P. Badin, G. Bailly, M. Raybaudi, and C. Segebarth, "A Three-dimensional Linear Articulatory Model Based on MRI Data," in *Proc. ETRW*, 1998.

[11] O. Engwall, "Combining MRI, EMA and EPG Measurements in a Three-dimensional Tongue Model," *Speech Communication*, vol. 41, no. 2-3, pp. 303–329, 2003.

[12] P. Birkholz, D. Jackel, and B. Kroger, "Construction And Control Of A Three-Dimensional Vocal Tract Model," in *Proc. ICASSP*, 2006.

[13] Y. Payan and P. Perrier, "Synthesis of VV Sequences with a 2D Biomechanical Tongue Model Controlled by the Equilibrium Point Hypothesis," *Speech communication*, vol. 22, no. 2-3, pp. 185–205, 1997.

[14] J.-M. Gérard, R. Wilhelms-Tricarico, P. Perrier, and Y. Payan, "A 3D Dynamical Biomechanical Tongue Model to Study Speech Motor Control," *arXiv preprint physics/0606148*, 2006.

[15] S. Fagel and K. Madany, "A 3-D Virtual Head as a Tool for Speech Therapy for Children," in *Proc. Interspeech*, 2008.

[16] P. Badin, Y. Tarabalka, F. Elisei, and G. Bailly, "Can You 'read' Tongue Movements? Evaluation of the Contribution of Tongue Display to Speech Understanding," *Speech Communication*, vol. 52, no. 6, pp. 493–503, 2010.

[17] A. Rathinavelu, H. Thiagarajan, and A. Rajkumar, "Three Dimensional Articulator Model for Speech Acquisition by Children with Hearing Loss," in *Proc. UAHCI*, 2007.

[18] H.-M. Wang, B. Chen, J.-W. Kuo, and S.-S. Cheng, "MATBN: A Mandarin Chinese Broadcast News Corpus," *Computational Linguistics and Chinese Language Processing*, vol. 10, no. 2, pp. 219–236, 2005.

[19] 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.

[20] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme Recognition Using Time-delay Neural Networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[21] V. Peddinti, D. Povey, and S. Khudanpur, "A Time Delay Neural Network Architecture for Efficient Modeling of Long Temporal Contexts," in *Proc. Interspeech*, 2015.

[22] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[23] R. Duan, T. Kawahara, M. Dantsuji, and H. Nanjo, "Efficient Learning of Articulatory Models Based on Multi-Label Training and Label Correction for Pronunciation Learning," in *Proc. ICASSP*, 2018.

[24] C. Zhang, Y. Liu, and C.-H. Lee, "Detection-based Accented Speech Recognition Using Articulatory Features," in *Proc. ASRU*, 2011.

[25] J.-C. Chen, J.-S. R. Jang, and T.-L. Tsai, "Automatic Pronunciation Assessment for Mandarin Chinese : Approaches and System Overview," *Computational Linguistics and Chinese Language Processing*, vol. 12, no. 4, pp. 443–458, 2007.

[26] W. Li, S. M. Siniscalchi, N. F. Chen, and C. H. Lee, "Improving Non-native Mispronunciation Detection and Enriching Diagnostic Feedback with DNN-based Speech Attribute Modeling," in *Proc. ICASSP*, 2016.

[27] D. L. Silver, "The Parallel Transfer of Task Knowledge Using Dynamic Learning Rates Based on a Measure of Relatedness," *Connection Science*, vol. 8, no. 2, pp. 277–294, 2002.

[28] T. Evgeniou and M. Pontil, "Regularized Multi-task Learning," in *Proc. ACM SIGKDD*, 2004.