A Study on Mispronunciation Detection Based on Fine-grained Speech Attribute

Minghao Guo*, Cai Rui*, Wei Wang*, Binghuai Lin†, Jinsong Zhang*, Yanlu Xie*
*Beijing Advanced Innovation Center for Language Resources, Beijing Language and Culture University, Beijing, China
E-mail: gmhgmh8000@163.com, cairui_blcu@163.com, wangwei_xinjiang@126.com, jinsong.zhang@blcu.edu.cn, xieyanlu@blcu.edu.cn Tel: +86-10-18911412480
† MIG, Tencent Science and Technology Ltd., Beijing, China
E-mail: binghuailin@tencent.com

***Abstract*—Over the last decade, several studies have investigated speech attribute detection (SAD) for improving computer assisted pronunciation training (CAPT) systems. The predefined speech attribute categories either is IPA or language dependent categories, which is difficult to handle multiple languages mispronunciation detection. In this paper, we propose a fine-grained speech attribute (FSA) modeling method, which defines types of Chinese speech attribute by combining Chinese phonetics with the international phonetic alphabet (IPA). To verify FSA, a large scale Chinese corpus was used to train Time-delay neural networks (TDNN) based on speech attribute models, and tested on Russian learner data set. Experimental results showed that all FSA's accuracy on Chinese test set is about 95% on average, and the diagnosis accuracy of the FSA-based mispronunciation detection achieved a 2.2% improvement compared to that of segment-based baseline system. Besides, as the FSA is theoretically capable of modeling language-universal speech attributes, we also tested the trained FSA-based method on native English corpus, which achieved about 50% accuracy rate.**

I. INTRODUCTION

The computer-aided pronunciation training (CAPT) system based on automatic speech attribute transcription (ASAT)[1], unlike traditional GOP-based CAPT system, takes speech attribute detection (SAD) as a front-end task to integrating phonetic knowledge (e.g., voicing and aspiration). Then, the SAD is used to improve two key functions in a CAPT system, namely detecting mispronunciation and providing feedback information [2-3]. In [4], the ASAT-based CAPT system obtained higher diagnostic accuracy of mispronunciation detection than the traditional phoneme-based GOP[5] method on automatic speech recognition (ASR). Furthermore, mispronunciation detection at a sub-segmental level, such as manner and place of articulation, can more accurately specify systematic pronunciation errors of second language (L2) [6]. However, most existing CAPT systems are often phoneme-based and SAD-based method are rarely used. Two of the most important reasons is lacking of large-scale training resources with qualified annotations, and how to accurately handle different language backgrounds SAD.

Several researches have explored the above-mentioned issue. [7] used additional L2 data set as a training set, which improve the robustness of SAD. As L2 pronunciation is easily affected by learners' native language, [8] proposed a SAD method which adaptively model speech attribute of L2 speakers by using their native language. For this method, the ASAT have demonstrated a potential of utilizing results from multi-language SAD as a bank of universal detectors [9]. This inference has been verified in multiple tasks[10-11]. In fact, a SAD system modeling all learners' native languages is difficult to implement. It is interesting to note that because each language has own unique dependencies between speech attributes, the language-discrimination of SAD will be more obvious when more precisely integrated attributes [12]. Therefore, we propose a fine-grained speech attribute (FSA) modeling method. Speech attributes have different modalities, e.g., the manner of articulation has discrete values while the position of the tongue has continuous values. By accurately describing values of Chinese speech attribute, the FSA method is more suitable for Chinese pronunciation habits.

In this paper, the FSA method defines seven types of speech attribute by referring to the definition of Chinese phonetics and mapping the Chinese phonemes to the international phonetic alphabet (IPA). In addition, the FSA method discretize the continuous values of Chinese final attribute into different dimensions, in which 5-dimensional and 7-dimensional values are respectively adopted to represent the tongue position (as shown in Table 2). Then, based on the ASAT paradigm, a bank of speech detectors is first built using ASR techniques to get information about the presence of speech attributes in 300 hours of Chinese corpus. In order to prevent the detectors from learning dependencies between different speech attributes, we modeled seven types of speech attribute separately.

As far as the SAD-based method, most of prior works focused on multi-language large vocabulary continuous speech recognition (LVCSR) [13-14] and low-Resource Speech Recognition [15]. In this work, we compared context-dependent and context-independent attribute based methods to observe the language-dependence of SAD. Moreover, we used MFCC and i-Vector [16] features to jointly model SAD-based Time-Delay Neural Networks (TDNN), then the model performance was then evaluated on Chinese and English corpus by the frame level recognition accuracy. Finally, we compared the diagnostic accuracy based on FSA method with the segment-based method on the Russian L2 learner data-set.

II. DEFINITION OF FINE-GRAINED SPEECH ATTRIBUTES

The attributes of speech include a set of fundamental speech sounds and their linguistic interpretation, a speaker profile encompassing gender, accent, etc [17]. In this paper, Chinese consonant sounds are described with four types of knowledge sources: place of articulation (PA), manner of articulation (MA), aspiration (AS) and voicing (VO). Chinese vowels include five attributes: tongue front-end (TF), tongue height (TH), rounding (RO), AS and VO. It is remarkable that speech attributes of consonants and vowels have been defined differently in acoustic phonetics[18], so we modeled the speech attributes of the vowels and consonants separately and tried to merge them in the PA classification. Since all Chinese vowels have no subcategories in the AS and VO, they are shown in the first part.

A. Definition of consonant speech attributes

We mapped all Chinese consonants with IPA one by one, and found the classification information we needed based on the phonetic knowledge of the corresponding phonemes on the IPA. In the MA, all vowel parts will be given the label "vowels", and four attributes were derived from the phone transcriptions using mapping tables (Table1 [19]). Chinese consonants represented by Pinyin are presented firstly followed by the English consonants given by Timit phoneme labels. The table also lists the attributes that exist in English but do not exist in Chinese, which are not modeled. It is not difficult to find the difference between Chinese and English attributes from the table. For example, there is no AS in English and only one unvoiced vowel "axh" exits in the phoneme set of Timit.

Tab. 1 Consonant attributes categories list

	Attributes	Phone set (Ch/En)	
P A	Bilabial	b p m	p b m w
	Labiodental	f	f v
	Alveolar	d t l n t l el ch sh jh zh dx nx	
	Dental	c s z	s dh en n r z th d
	Retroflex	zh ch sh r	
	Palatal	j q x	y
	Velar	g k h	k g ng
M A	Stop	g p d t g k	t p k b d g
	Fricative	f s sh r x h	sh th f hh dh hv v w zh s z
	Affricate	z zh c ch j q	ch jh
	Nasal	m n	en m nx ng n
	Lateral	l	el l
	Approximant		dx
	Tap or Flap		r y
A S	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 vowels	
V O	Voiced	l m n r vowels b dh dx d el en g jh hv l m nx ng n r v w y zh z	
	Unvoiced	b c ch d f g h j k p q s sh t x z zh Other_vowels ch sh s th t f p hh k axh	

B. Definition of vowel speech attributes

Chinese finals composed of multiple vowels and nasal vowels(e.g., “en”, “iang”, etc.) are relatively complex compared to the initials and attributes of Chinese finals are continuous values. Therefore, we discretize these continuous values into different dimensions, subsequently get the set of speech attribute of each final based on IPA phonemes. What deserves our attention is that there are three types of attribute set in the Chinese finals, which describe how many dimensions exist in this final. For example, the Chinese final "iao" was described as three IPA phonemes, so it is three dimensions in each attribute. In Table 2, four other Chinese vowel attributes are shown. As the complexity of multi-dimensional finals, only all dimensions of each attribute and corresponding Uni-dimensional finals are listed.

In addition, the vowels in Chinese and English differ greatly in tongue position(TP). In the past works, the TP were roughly defined as three dimensions: front, middle and back [20]. In order to describe the Chinese vowels more accurately, we divided TP into five dimensions and seven dimensions respectively according to the definition of Chinese phonetics and IPA. Since the five-dimensions TP can directly correspond to the initials, we try to model the finals and initials simultaneously in the PA category. There are more detailed seven-dimensions TP in the TF category. Moreover, the Chinese initial is marked as "consonants" in TF, RO and TH.

Tab. 2 Vowel attributes categories list

	Attributes	Phone set (Ch/En)	
P A	Dental	ii	
	Retroflex	iii	
	Palatal/Front	i v	iy ih ae eh
	PA-Central	a	ax ix ux axh axr er
	Velar/Back	u	aa ah ao uw uh
T H	High	i ii iii v u	ix iy ux uw
	Second H		ih uh
	Half H		
	Middle		axh axr ax
	Half L		ah ao eh er
	Second L		ae
	Low	a	aa
T F	Front 2	ii	
	Front 1	iii	
	Front	i v	ae eh iy
	Half F		ih
	Central	a	axh axr ax er ix ux
	Half B		uh uw
	Back	u	aa ah ao
R O	Rounded	u v	ao uw ux
	Unrounded	a i ii iii	aa ae ah ax eh er ih ix iy uh axr axh

III. MISPRONUNCIATION DETECTION BASED ON FSA MODELING METHOD

The frame-level attribute features can be used to formulate linguistic knowledge for pronunciation changes caused by either regional accent or co-articulation as context-dependent rules associated with substitutions of different features. In this work, we obtained the corresponding frame-level attribute feature after modeling the above seven attribute detectors. The FSA-based mispronunciation detection framework is shown in Figure 1.

A. FSA-based modeling

HMM/TDNN framework is used to design attribute detectors in the proposed approach, and Time-Delay Neural Networks (TDNN) have been shown to be a good method for the classification of dynamic speech sounds such as voiced stop consonants[21]. Moreover, HMM-based ASR typically model each phoneme using 3 states (begin, middle, end) to account for co-articulation, and previous work has shown that using only the middle frames for training speech attribute detectors leads to the best results[22]. In the FSA-based modeling process, the context-dependent HMM may be over-fitting for the Chinese pronunciation habit and the TDNN already has the ability to simulate the long-term dependence of the speech attribute. Therefore we compared the performance of three modeling methods, namely Monophone HMM+TDNN, Triphone HMM+TDNN, Monophone HMM+Context-independent DNN.

In the field of ASR, there are methods like i-Vectors to adapt neural networks to different speakers, and these methods show that neural networks benefit from additional input features. Hence, i-Vector features are used to distinguish speaker information and train deep neural networks with MFCC, which can eliminate interference of speaker information in universal-attribute detection tasks at the feature level.

According to above the FSA method, we established seven TDNN-based attribute classifiers. However, modeling the initials and finals of Chinese respectively led to an unbalanced distribution of training data. For example, the useless initial label "consonants" in the TH category has nearly half of the training data. The phone-based background model (PBM) is adopted to address this issue. The key idea is to generate a multiple-label representation of the useless class, which can be achieved by dividing N/A classes into several sub-classes.

B. FSA-based detection framework

The front-end feature extraction module consists of a bank of speech attribute classifiers, and mispronunciation detection can be defined on various time-scales, namely supra-segmental (e.g., lexical stress), segmental (e.g., substitution of phonetic units), and sub-segmental (e.g., voicing feature activated for a canonical unvoiced phone [23,24]). Expanded frames of input speech (MFCC) and i-Vector features are fed into each front-end classifier, then the current frame likelihoods pertaining to each possible attribute within that category are generated. As shown in Figure1, a group of the frame attribute posteriors was used to evaluate the cross-language ability of FSA modeling methods on Chinese and English test sets, and feed into the back-end module for segmental mispronunciation detection, and generate phoneme level posterior probability for sub-segmental mispronunciation detection. Moreover, we have completed pronunciation error detection on the Russian L2 learner corpus, in which the phoneme boundary information of the audio is also obtained by another independent force alignment. Equation (1) is used to calculate phone level log posterior by force-alignment [25]:

$$\log P(p \mid o; t_s, t_e) = \frac{1}{t_e - t_s} \sum_{t_s}^{t_e} \log \sum_{s \in p} P(s \mid o_t), \tag{1}$$

Where o_t is the input feature at frame t; t_s and t_e are the start and end time of unit p, obtained by forced alignment or annotation information from Timit. $P(s|o_t)$ is frame level likelihood; $\{s \in p\}$ is the set of context-dependent or context-independent units, whose central unit is p.

IV. ATTRIBUTE RECOGNITION EXPERIMENT

A. Speech corpora

The training speech corpus is from the Chinese National Hi-Tech Project 863 for Mandarin LVCSR system development [26], and the Aishell 178 hours Mandarin corpus on Kaldi toolkit. A total of 250,000 utterances spoken by 1800 speakers (300 hours) were used for acoustic modeling. Sufficient data ensures the robustness based on FSA method modeling. There are two test sets, one is 7000 utterances of Chinese data from Aishell corpus and the other is 6000 utterances of English data from Timit. Chinese L2 speech database can be referred to BLCU inter-Chinese speech corpus[27], containing 5000 utterances spoken by 48 Russian learners of Mandarin.

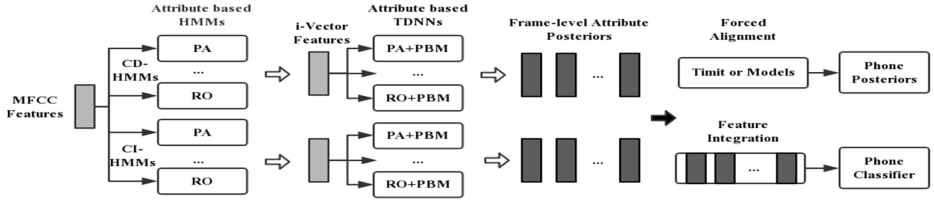

Figure 1: the FSA-based mispronunciation detection framework

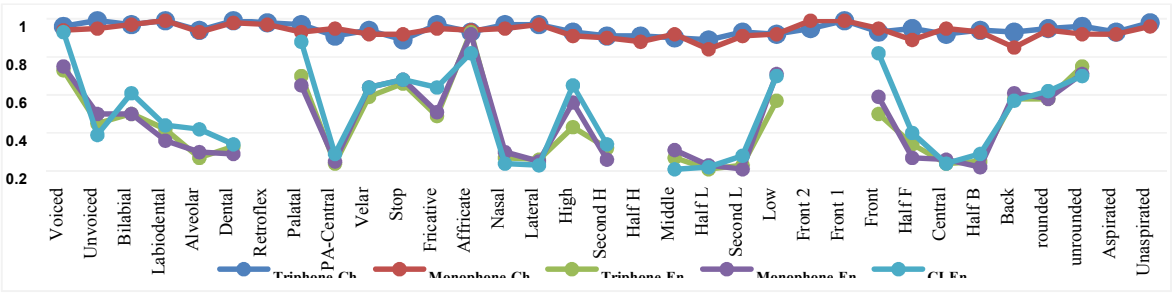

Figure 2: FSA based detectors accuracy on Chinese test set and English test set.

B. *Experimental results*

In the following sections, native-language (Ch) and cross-language (En) speech attribute detection are evaluated, and the experimental results of all attributes from there modeling methods (Triphone HMM+TDNN, Monophone HMM+TDNN, Monophone HMM+Context-independent DNN) are shown in Figure 2. The top two curves show reliable performance on the native-language test set, and the below three curves demonstrate that relatively low attribute accuracy are achieved on across-languages test set, especially the vowel part. This can be explained by thinking of the complicated structure of the English vowels. On the Chinese test set, the performance of the two methods is comparable, but on the English test set, the performance as the whole is better when the model has less Chinese context information. This phenomenon reflects the linguistic independence of universal speech attributes. A more in depth analysis reveals that several attributes, such as Affricate (93%) and Voiced (78%), can achieve good attribute accuracy on the English test set. Subsequently, we find that attribute accuracy from TF with more refined classification is comparable with PA on En, which shows that TF can better adapt to cross-language speech attributes.

V. PRONUNCIATION ERROR DETECTION

In order to detect pronunciation errors on sub-segmental and segmental, F-score and diagnostic accuracy (DA) are used to evaluate the performance of each mispronunciation system.

$$DA = \frac{N_M + N_C}{N} * 100\ \% \quad (1)$$

$$Precision = \frac{N_M}{N_D} * 100\ \% \quad (2)$$

$$Recall = \frac{N_M}{N_E} * 100\ \% \quad (3)$$

$$F\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

where N_M is the number of true mispronunciations detect and the detection results are consistent with the human annotations. N_C is the number of true correct pronunciation detected by the system. N_D is the number of all detected pronunciation errors. N_E is the total number of pronunciation errors in the test set. N is the number of phone or attribute in the test set. Table 3 shows the sub-segmental mispronunciation detection performance. Table 4 compares two systems: the FSA-based and segmental-based systems with the same training set at segmental level. We selected seven classifiers with better detection performance at sub-segmental level to evaluate the pronunciation quality of second language learners and listed in table 3. We can see that FSA-based methods can generalize well in different speech attributes. The performance of TF is better than PA, which indicates that the refinement of speech attribute categories is beneficial to mispronunciation detection. Compared with the segment-based pronunciation error detection, detection performance of FSA-based pronunciation error is higher.

Table 3: Diagnostic accuracy at the sub-segmental mispronunciation detection.

	VO	AS	MA	PA	TH	TF	RO
DA	89%	89%	87%	83%	86%	84%	88%

Table 4: DA and F-score at the segmental mispronunciation.

	FSA-based	Segment-based
F-score	71.5%	63.5%
DA	86.5%	84.3%

VI. CONCLUSIONS

In this paper, we proposed a modeling method based on the fine-grained speech attribute(FSA) on Chinese corpus to detect mispronunciation. Experimental results have shown that this approach reliably extracts frame-level accuracy rate of speech attributes and achieves better detection results than segment-based approaches. By comparing TF and PA, the benefits of accurately describing speech attributes are also demonstrated. On the English corpus, experimental results show the FSA can be used in any language theoretically.

ACKNOWLEDGMENT

This study was supported by Advanced Innovation Center for Language Resource and Intelligence (KYR17005), the Fundamental Research Funds for the Central Universities（18YJ030004）, National social Science foundation of China (18BYY124), BLCU support project for young researchers program (19YCX131) (the Fundamental Research Funds for the Central Universities), and, "Intelligent Speech technology International Exchange" Introduced Intelligence Project. Yanlu Xie is the corresponding author.

REFERENCES

[1] C.-H. Lee, "From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition," in Proc. Interspeech, Jeju Island, Korea, Oct. 2004, pp. 109–112.

[2] Richeng Duan, Jingsong Zhang, Wen Cao, Yanlu Xie, "A preliminary study on asr-based detection of Chinese mispronunciation by Japanese learners", INTERSPEECH 2014.

[3] Yingming Gao, Yanlu Xie, Wen Cao, Jinsong Zhang, "A Study on Robust Detection of Pronunciation Erroneous Tendency Based on Deep Neural Network", INTERSPEECH 2015.

[4] Li W . Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling[C]// IEEE International Conference on Acoustics. IEEE, 2016.

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

[6] 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.

[7] Li, Wei, et al. "Improving Mispronunciation Detection for Non-Native Learners with Multisource Information and LSTM-Based Deep Models." INTERSPEECH. 2017.

[8] Duan R , Kawahara T , Dantsuji M , et al. Effective articulatory modeling for pronunciation error detection of L2 learner without non-native training data[C]// ICASSP 2017 - 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

[9] Chiang C Y , Siniscalchi S M , Wang Y R , et al. A study on cross-language knowledge integration in Mandarin LVCSR[C]// International Symposium on Chinese Spoken Language Processing. IEEE, 2012

[10] S. Stüker, T. Schultz, F. Metze, and A. Waibel, "Multilingual articulatory features," in Proc. ICASSP, Hong Kong, Hong Kong, Apr. 2003,pp. 144–147.

[11] S. M. Siniscalchi and C.-H. Lee, "A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition,"Speech Commun., vol. 51, pp. 1139–1153, 2009.

[12] Siniscalchi S M , Lyu D C , Svendsen T , et al. Experiments on Cross-Language Attribute Detection and Phone Recognition With Minimal Target-Specific Training Data[J]. IEEE Transactions on Audio, Speech and Language Processing, 2012, 20(3):875-887.

[13] J. Köhler, "Multilingual phoneme recognition exploiting acoustic-phonetic similarities of sounds," in Proc. ICSLP, Philadelphia, PA, Oct.1996.

[14] S. Goksen and J. N. Gokcen, "A multilingual phoneme and model set: Towards a universal base for automatic speech recognition," in Proc. IEEE Workshop Automatic Speech Recognition and Understanding, Santa Barbara, CA, Dec. 1997, pp. 599–605.

[15] MarkusM¨uller, Sebastian St¨uker, and AlexWaibel, "Towards Improving Low-Resource Speech Recognition Using Articulatory and Language Features," in Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT), Seattle, U.S.A., 2016.

[16] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, Front-end factor analysis for speaker verification IEEE Transactions on Audio Speech & Language Processing, vol. 19, no.4, pp. 788-798, 2011

[17] Siniscalchi S M , Lee C H . An attribute detection based approach to automatic speech processing[J]. 2014.

[18] S. M. Siniscalchi et al, "Toward a detector-based universal phone recognizer," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Las Vegas, Nevada, USA, Mar./Apr. 2008, pp. 4261–426

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

[20] Muller M , Franke J , Waibel A , et al. Towards phoneme inventory discovery for documentation of unwritten languages[C]// ICASSP 2017 - 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

[21] Florian Metze, Articulatory Features for Conversational Speech Recognition, Ph.D. thesis, Karlsruhe, Univ., Diss., 2005, 2005

[22] Hou J , Rabiner L R , Dusan S . On the use of time-delay neural networks for highly accurate classification of stop consonants[C]// INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007. DBLP, 2007.

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

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

[25] 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.

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

[27] 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.