EFFICIENT LEARNING OF ARTICULATORY MODELS BASED ON MULTI-LABEL TRAINING AND LABEL CORRECTION FOR PRONUNCIATION LEARNING

Richeng Duan[1], Tatsuya Kawahara[1], Masatake Dantsuji[2], Hiroaki Nanjo[2],

[1]School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
[2]Academic Center for Computing and Media Studies, Kyoto University, Japan

ABSTRACT

Articulatory feedback is effective for computer-assisted pronunciation training (CAPT) systems. This paper investigates efficient model learning methods for providing articulatory information to language learners. We first propose an articulatory attribute modeling method based on a multi-label learning scheme. Then, the models are further enhanced with a simple and effective training label correction method. These proposed methods are evaluated in three tasks: native attribute recognition, pronunciation error detection of non-native speech, and non-native speech recognition. Experimental results show that proposed methods significantly improve the conventional deep neural network (DNN) based articulatory models.

Index Terms—Computer-assisted pronunciation training (CAPT), pronunciation error detection, articulation modeling, multi-label DNN, attribute label correction

1. INTRODUCTION

Over the last decades, CAPT systems based on statistical modeling techniques have made considerable progress [1-5]. There are generally two kinds of pronunciation feedback provided in these systems. One is to show learners pronunciation scores [6-9], and the other detects individual errors such as phone substitution errors [10-14]. A typical scenario is: "You made an r-l substitution error." when a user pronounces the word "red" as "led". Instead of providing phone substitution feedback, giving the feedback directly related with articulation is more attractive in recent years [15-17]. Facing the same pronunciation error described above, learners could be instructed with "Try to retract your tongue and make the tip between the alveolar ridge and the hard palate". Articulatory information has been demonstrated more helpful in many related areas, such as speech comprehension improvement [18], speech therapy [19] and pronunciation perceptual training [20].

We aim at providing such articulation related feedback and have investigated modeling the articulatory attributes through transfer-learning methods [21]. In this work, we introduce two improvements. Firstly, we take care of the interaction effects of different kinds of articulation attributes of the same phone with a multi-label training scheme. A DNN model with multiple outputs is designed, in which the hidden layer can be regarded as a shared internal feature representation of vocal tract configuration. Another benefit of this learning method is that it can largely reduce the model training time compared with the conventional attribute modeling methods. In addition, the models are further improved with a label correction procedure based on the consistency of the articulation attribute labels.

The rest of this paper is organized as follows: Conventional articulatory attribute modeling methods are firstly described in Section 2. All details of proposed training methods are explained in Section 3 and 4. Section 5, 6 and 7 respectively report the performance of these learning methods in the native attribute recognition task, the non-native pronunciation error detection task, and non-native speech recognition task. Conclusions are in the final section.

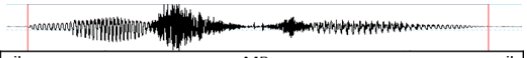
2. ARTICULATORY MODELING

Articulation means the movement of the tongue, lips, and other organs to generate speech sounds. Generally, place of articulation 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). We have investigated articulatory models to recognize these attributes.

2.1. Articulatory attributes transcription

For the training of the articulatory attributes with supervised statistical models, high-quality articulatory datasets are needed, which contain accurate articulatory position information along with speech recordings. Various methods are used to generate a speaker's articulatory attributes, including X-rays [22], electromagnetic articulography (EMA) [23], magnetic resonance imaging (MRI) [24] and ultrasounds [25]. However, all of the above direct measurements have their own disadvantages [26]. Moreover, it is not easy to obtain such dataset in a large scale. Therefore, the attribute transcriptions in present work are

derived from the phone transcription according to the phone-to-attribute mapping rules, which is a practical option adopted by many researchers [27-29]. From the example in Fig. 1, we can see the mapping relation between the phone class and the attribute class is many-to-many (phone /M/ has two attributes nasal and bilabial while both vowels /IH/ and /AX/ are mapped to the unrounded attribute). As a result, we prepare four kinds of articulatory transcriptions (manner, place-roundedness, place-backness and place-height) to represent all attributes. In each kind of transcription, the attributes are disjoint to each other so that it can be used to train a DNN model.



Sentence	sil	MR.						sil
Phone	sil	M	IH	S	T	AX	R	sil
Manner (1)	sil	nasal	vowel	unvoiced-fricative	unvoiced-stop	vowel	approximant	sil
Place & Backness (2)	sil	bilabial	front	alveolar	flap	central	palato-alveolar	sil
Place & Height (3)	sil	bilabial	mid	alveolar	flap	mid	palato-alveolar	sil
Place & Roundedness (4)	sil	bilabial	unrounded	alveolar	flap	unrounded	palato-alveolar	sil

Fig. 1. Converting phone labels to articulatory labels.

2.2. DNN based articulatory attribute modeling using phone-to-attribute transcription

Following the great success of DNN based acoustic modeling, articulatory attribute modeling with DNN has been investigated in recent years. According to the above-defined phone-to-attribute mapping, researchers usually train a bank of DNNs separately [27] [29]. The number of models depends on the attribute classes. In our previous work, we trained four DNNs in which each DNN was used to represent one-kind attribute. The co-articulation effects can be partially considered with context-dependent tri-attribute units. Similar to tri-phones used in ASR, labels for tri-manners and tri-places are generated by taking into account the labels of neighboring attributes. However, it cannot take into account the interaction effects among different attribute categories. For example, during the training procedure of tri-manners, only the temporal effect of articulation manner is modeled. In other words, information contained in the placement of articulation is not used during this kind of isolated training procedure.

3. ARTICULATORY ATTRIBUTE MODELING WITH MULTI-LABEL LEARNING

Based on the discussion above, we propose a new training method to the attribute modeling, which considers all the interaction effects in a unified objective function. Fig. 2 illustrates an overview of multi-label learning based on DNN, which we refer as multi-label DNN (ML-DNN). All the isolated DNNs (four in our work) are merged into a single ML-DNN. The objective function of this multi-label learning is a summation of four cross entropy loss:

$$Loss_{mtl} = CE_{man} + CE_{pl-bk} + CE_{pl-ht} + CE_{pl-rd}$$

This architecture shares parameters and updates them with multiple complementary labels. Note that except for more output nodes, the number of parameters in this learning scheme is same as that of isolated DNNs. Therefore, the training time would be roughly reduced by a factor of N, where N is the number of attribute categories.

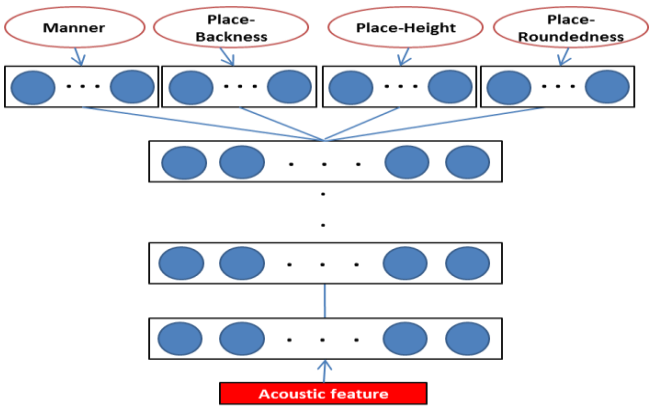


Fig. 2. Articulatory attribute modeling with ML-DNN.

4. CORRECTING TRAINING LABELS WITH VOTING PROCESS

Before training the DNN articulatory models, we usually generate the frame level training labels from the forced alignment process of a seed model. These generated labels may not be accurate enough, but all of them are used as ground-truth labels in the DNN training process. The different kinds of labels should be mapped to the same phone based on the linguistic mapping rules mentioned in Section 2.1. Taking the sentence in Fig. 1 for an example, if a speech frame belongs to the phone /M/, its attribute labels after forced alignment should get mapped to the nasal attribute in manner transcription (1) and the bilabial attribute in the placement transcriptions (2-4).

Motivated by the above discussion and the multi-label learning scheme, we propose a method to automatically correct noisy training labels through voting. Initially, four independent models are trained with the sentence level attribute transcription. The alignment outputs from these models are then sent to a correction module, which reconciles differences among these labels based on a voting scheme. The simple majority voting approach is adopted in present work. Since there are totally four attribute categories used in current classification, a label may only be changed when the other three attribute labels mapped to a same phone. The label correction procedure for a speech frame labeled "Unvoiced-fricative Alveolar Alveolar Unrounded" is shown in Fig. 3. These labels will be processed along the direction that is represented by red dashed arrows. After correction, the "Unrounded" label will be changed to its right neighboring attribute "Alveolar".

The procedure is iterated. It is not guaranteed to convergence in theory, but the number of corrections is decreased empirically. The procedure will be terminated when the proportion of corrected labels is smaller than a threshold (1e-5 in this work). Finally, the module will output newly corrected transcriptions for the model training. We use only attribute models, not phone models, because the recognition performance of phone models, especially in non-native speech, is much lower than that of attribute recognition.

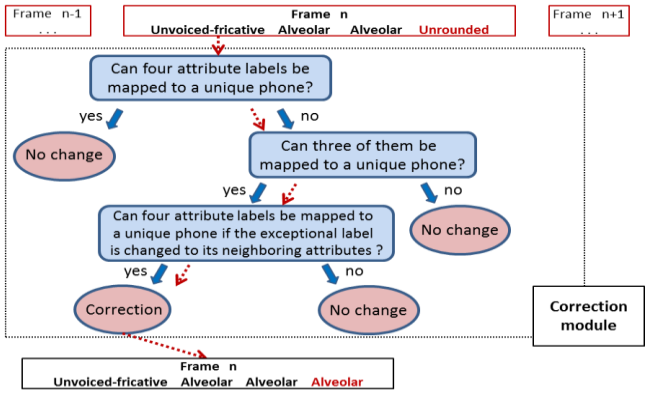


Fig. 3. Majority voting based training labels correction.

5. NATIVE ATTRIBUTE RECOGNITION EXPERIMENT

5.1. Data set and model configuration

The language learners in this study are Japanese students who learn English. As a consequence, native English and Japanese English are considered in this paper. We first evaluate our proposed methods on a native English corpus in this section. The corpus used to train the articulatory models is Wall Street Journal (WSJ) database [30], which is a commonly used database for English large-vocabulary continuous speech recognition research. Sixty-four hour speech data from the SI-284 training data (WSJ0 and WSJ1) were selected after filtering noisy utterances. We conduct the evaluation on both Nov'92 and Nov'93 testing data sets of WSJ.

All methods used the following DNN configuration, which is optimized on the standard development data set (Dev'93) of WSJ. The acoustic feature consists of 40-dimensional log Mel-scale filterbank outputs plus first and second temporal derivatives. The input to the network is 11 frames, 5 frames on each side of the current frame. The neural network has 7 hidden layers with 2048 nodes per layer. DNN training consists of unsupervised pre-training and supervised fine-tuning.

5.2. Effects of multi-label learning

The recognition results of English native articulatory attributes are shown in Table 1. Compared to baseline DNN models which are separately trained as conventionally done [21], ML-DNN improves the recognition performance of all attribute classes. From the last row of Table 1, we see that ML-DNN significantly reduced the error rate on both testing data sets. The statistical significance was confirmed in a two-sided t-test at a significance level of 0.05. The relative improvement column shows the average improvement over two testing datasets. We observe a large effect in "Place-Backness" and "Place-Height". It suggests that sharing the network by place-of-articulation attributes is effective. Note also that the absolute performance of the two attributes was lower than others, so they had much potential of improvement.

Table 1. Effect of multi-label learning in native attribute recognition (error rate %).

Attribute	Dataset	DNN	ML-DNN	Relative Improvement
Manner	Nov'92	6.32	6.14	1.72
	Nov'93	8.32	8.27	
Place-Roundedness	Nov'92	8.60	8.41	1.93
	Nov'93	10.27	10.10	
Place-Backness	Nov'92	10.70	9.64	9.20
	Nov'93	13.07	11.96	
Place-Height	Nov'92	9.75	9.08	5.24
	Nov'93	12.46	12.01	
Overall	**Nov'92**	**8.85**	**8.33**	
	Nov'93	**11.01**	**10.60**	

5.3. Effects of training label correction

We show the effects of correcting label method in Table 2. We see that models trained with corrected labels yielded a lower error rates in both testing sets. The error rate is further reduced to 8.03% in Nov'92. The relative improvement is about 10% compared to the DNN baseline (8.85%). The same tendency is observed in the Nov'93 dataset. We also observe similar relative improvements among different attributes, which suggests that label correction was applied almost uniformly to all attributes. It is suggested that the inconsistencies among different categories are random, and they can be corrected by the proposed method.

Table 2. Effect of label correction in native attribute recognition (error rate %).

Attribute	Dataset	ML-DNN	ML-DNN + Correction	Relative Improvement
Manner	Nov'92	6.14	6.06	3.19
	Nov'93	8.27	7.85	
Place-Roundedness	Nov'92	8.41	8.17	3.21
	Nov'93	10.10	9.74	
Place-Backness	Nov'92	9.64	9.25	3.32
	Nov'93	11.96	11.65	
Place-Height	Nov'92	9.08	8.93	3.45
	Nov'93	12.01	11.38	
Overall	**Nov'92**	**8.33**	**8.03**	
	Nov'93	**10.60**	**10.10**	

6. PRONUNCIATION ERROR DETECTION OF LANGUAGE LEARNERS

In this section, we apply the proposed methods to non-native speech for pronunciation error detection. We detect the pronunciation errors directly on the attribute level, which is different from the phone level error detection conducted in other works [10-14].

6.1. Experimental settings

The evaluation data is a corpus of English words spoken by Japanese students [31]. There are 7 speakers (2 male, 5 female) and each speaker uttered a same set of 850 English words. Pronunciation errors of vowels are focused in this experiment. We employ finite state decoding network [13] for pronunciation error detection, which includes the canonical pronunciation and possible pronunciation errors. Detection accuracy (DA) [15-16] is used to evaluate the performance of different methods.

6.2. Experimental results

Fig. 4 compares the overall pronunciation error detection performance of three different methods: conventional DNN, ML-DNN, and ML-DNN trained with corrected labels (ML-DNN + correction). We observe the effects of proposed methods when they are applied to the pronunciation error detection of non-native speech. Compared to the conventional DNN, ML-DNN improves DA significantly (0.05 significance level) from 72.10% to 74.00%. Similar to the native attribute recognition task, label correction method further improved the performance of pronunciation error detection. However, the absolute performance is much lower than that of native attribute recognition. This is mainly caused by the mismatch that attribute models were trained only using native speech while testing samples are non-native. We address this problem by incorporating a native speech database of the language learners [21].

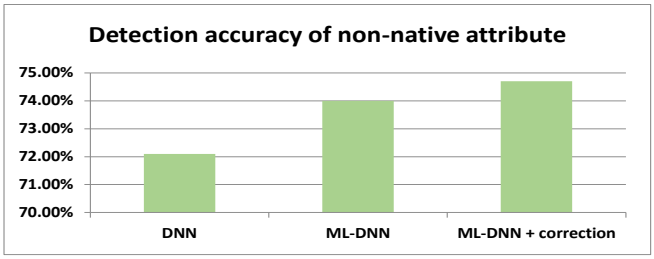


Fig. 4. Overall detection accuracy of different methods.

7. NON-NATIVE WORD RECOGNITION OF LANGUAGE LEARNERS

7.1. Task and recognition configuration

Finally, the proposed methods are evaluated on non-native speech recognition task, which is a necessary module in advanced conversation based pronunciation learning system. Similar to that used in [28], the attribute classification is used as the secondary task to improve the speech recognition performance. We conducted word recognition experiments with different settings. One is continuous speech recognition (CSR) while the other is isolated word recognition (IWR) which is more constrained. The evaluation data is same as what we used in the previous section. Considering the pronunciation variation of non-native speech, we also conduct experiments with an extended lexicon [32], in which each word is represented by both canonical pronunciation and other possible pronunciation variations. These added pronunciations are derived based on the study of phonological properties of the native language and the target language.

7.2. Experimental results

From Fig. 5, we see that proposed methods consistently perform better in all different recognition settings. ML-DNN trained with corrected labels achieved lower WER, which are 3.57%, 2.11%, 2.34% and 2.06% absolute improvement from the conventional DNN method. All these improvements are significant at the significance level of 0.05 in two-sided t-test.

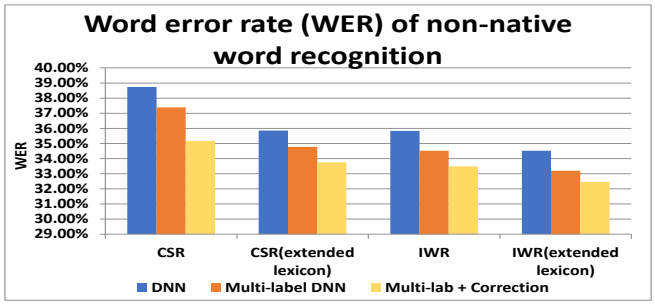


Fig. 5. Non-native English word recognition.

8. CONCLUSIONS

In this paper, we present two methods for efficiently learning the articulatory models. The multi-label learning method allows for learning all the attributes at the same time with parameter sharing. This single model with multiple outputs can also model the interaction effects of different articulators. Moreover, it can largely reduce the training time than the separate training scheme. The correcting method can enhance the attribute models with corrected frame level training labels. It can reconcile differences among different attribute labels through a voting process.

There are several directions for future work: one is to construct a hierarchical acoustic phone model by adding a phone classification layer on top of this ML-DNN. We will extend the current simple voting method to some confidence score based voting methods. Using soft corrected labels may also be more effective than the present hard corrected labels.

REFERENCES

[1] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy technology interface in computer assisted pronunciation training," in Computer assisted language learning, 2002.

[2] S. Seneff, C. Wang, and J. Zhang, "Spoken conversational interaction for language learning," in Proceedings of the InSTIL/ICALL Symposiumon on Computer Assisted Learning, pp. 151-154, 2004.

[3] R. Downey, H. Farhady, R. Present-Thomas, M. Suzukiet, and M. Van, "Evaluation of the usefulness of the Versant for English Test: A response," in Language Assessment Quarterly, pp. 160-167, 2008.

[4] H. Strik, J. Colpaert, J. Doremalen, and C. Cucchiarini, "The DISCO ASR-based CALL system: practicing L2 oral skills and beyond," in Proceedings of International Conference on Language Resources and Evaluation. Istanbul, pp. 2702-2707, 2012.

[5] X. Qian, H. Meng, and F. Soong, "A Two-Pass Framework of Mispronunciation Detection and Diagnosis for Computer-Aided Pronunciation Training," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(6), pp.1020-1028, 2016.

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

[7] S.Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," in Speech Communication, vol. 30, pp. 95–108, 2000.

[8] F. Zhang, C. Huang, F.K. Soong, M. Chu, and R.H. Wang, "Automatic mispronunciation detection for Mandarin," in Proc. ICASSP, 2008.

[9] 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," in Speech Communication, vol 67, pp. 154–166, 2015.

[10] H. Strik, K. Truong, F. De Wet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," in Speech Communication, vol51, pp. 845-852, 2009.

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

[12] Y.-B. Wang and L.-S. Lee, "Toward unsupervised discovery of pronunciation error patterns using universal phoneme posteriorgram for computer-assisted language learning," in Proc. ICASSP, 2013.

[13] A. Lee and J. Glass, "Personalized mispronunciation detection and diagnosis based on unsupervised error pattern discovery," in Proc. ICASSP, 2016.

[14] S. Joshi, N. Deo, and P. Rao, "Vowel mispronunciation detection using DNN acoustic models with cross-lingual training," in Proc. Interspeech, 2015.

[15] 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 , pp. 6135-6139, 2016

[16] H. Ryu, M. Chung, "Mispronunciation Diagnosis of L2 English at Articulatory Level Using Articulatory Goodness-Of-Pronunciation Features", in Proc. SLATE , 2017

[17] R. Duan, J. Zhang, W. Cao, and Y. Xie. "A Preliminary study on ASR-based detection of Chinese mispronunciation by Japanese learners", in Proc. Interspeech, 2014.

[18] P. Badin, Y. Tarabalka, F. Elisei, and G. Bailly, "Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding," in Speech Communication, vol. 52, pp. 493-503, 2010.

[19] S. Fagel and K. Madany, "A 3D virtual head as a tool for speech therapy for children," in Proc. Interspeech, 2008.

[20] A. Rathinavelu, H. Thiagarajan, and A. Rajkumar, "Three dimensional articulator model for speech acquisition by children with hearing loss," in Proceedings of the 4th Internation Conference on Universal Access in Human Computer Interaction, vol. 4554, pp. 786-794, 2007.

[21] R.Duan, T.Kawahara, M.Dantsuji, and J.Zhang, "Effective articulatory modeling for pronunciation error detection of L2 learner without non-native training data", In Proc. ICASSP, pp.5815--5819, 2017.

[22] J.R. Westbury, "X-ray Microbeam Speech Production Database User's Handbook", Waisman Center on Mental Retardation and Human Development. University of Wisconsin, Madison, WI, USA, version 1.0 edition. 1994.

[23] AA. Wrench, "Multi-channel/multi-speaker articulatory database for continuous speech recognition research", Phonus 5, 1–13, 2000.

[24] S. Narayanan, K. Nayak, S. Lee, A. Sethy, D. Byrd, "An approach to real-time magnetic resonance imaging for speech production", J. Acoust. Soc. Am. 115, 1771–1776, 2004.

[25] M. Grimaldi, B.F. Gili, F. Sigona, M. Tavella, P. Fitzpatrick, L. Craighero, L. Fadiga, G. Sandini, and G. Metta, "New technologies for simultaneous acquisition of speech articulatory data: 3D articulograph, ultrasound and electroglottograph", In Proceedings LangTech, Italy, 2008.

[26] H. Li, J. Tao, M. Yang, B. Liu, "Estimate articulatory MRI series from acoustic signal using deep architecture", in Proc. ICASSP, 2015.

[27] B. Abraham, S. Umesh, "An automated technique to generate phone-to-articulatory label mapping", in Speech Communication, vol 86, pp. 107–120, 2017.

[28] H. Zheng, Z. Yang, L Qiao, J. Li, W. Liu, "Attribute Knowledge Integration for Speech Recognition Based on Multi-task Learning Neural Networks", in Proc. INTERSPEECH, 2015

[29] W. Li, S. Marco, Nancy 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.

[30] Douglas B Paul and Janet M Baker, "The design for the Wall Street Journal-based CSR corpus," in Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, 1992, pp. 357–362

[31] K. Tanaka, H. Kojima, Y. Tomiyama, and M. Dantsuji, Acoustic models of language independent phonetic code systems for speech processing. Spring Meeting of the Acoustical Society of Japan: Proceedings. Tokyo: Acoustical Society of Japan, 1:191–192, 2001.

[32] S. Schaden, "Generating non-native pronunciation lexicons by phonological rule," in Proc. ICSLP, pp.2545–2548, 2004