EFFECTIVE ARTICULATORY MODELING FOR PRONUNCIATION ERROR DETECTION OF L2 LEARNER WITHOUT NON-NATIVE TRAINING DATA

Richeng Duan[1], Tatsuya Kawahara[1], Masatake Dantsuji[2], Jinsong Zhang[3],

[1]School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
[2]Academic Center for Computing and Media Studies, Kyoto University, Japan
[3]School of Information Science, Beijing Language and Culture University, China

ABSTRACT

For effective articulatory feedback in computer-assisted pronunciation training (CAPT) systems, we address effective articulatory models of second language (L2) learners' speech without using such data, which is difficult to collect and annotate in a large scale. Context-dependent articulatory attributes (placement and manner of articulation) are modeled based on deep neural network (DNN). In order to efficiently train the non-native articulatory models, we exploit large speech corpora of native and target language to model inter-language phenomena. This multi-lingual learning is then combined with multi-task learning, which uses phone-classification as a sub-task. These methods are applied to Mandarin Chinese pronunciation learning by Japanese native speakers. Effects are confirmed in the native attribute classification and pronunciation error detection of non-native speech.

Index Terms—Computer-assisted pronunciation training (CAPT), pronunciation error detection, articulation modeling, DNN, multi-lingual learning

1. INTRODUCTION

With the accelerating process of globalization, there is an increasing need for learning a second language. CAPT systems provide opportunities for learners practising their pronunciation in a stress-free environment. Over the last decades, CAPT systems based on statistical modeling techniques have made considerable progress [1-8]. In general, there are two kinds of pronunciation feedback in the current systems. One is to give learners pronunciation scores [9-15], and the other detects individual errors such as specific phone substitution errors [16-25]. According to the scores, learners can know their pronunciation proficiency, but they cannot know what the errors are and how to correct them when getting a low score. Regarding detection of phone substitution errors, some researchers target a few specific problematic phones while others build systems with the automatic speech recognition (ASR) technology, which is more general than the specially designed ones. 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. 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". This approach has been demonstrated more helpful in many areas, such as speech comprehension improvement [26], speech therapy [27] and pronunciation perceptual training [28]. One direct way of achieving this goal is to train the articulatory models of L2 learners. However, it is not easy to collect a non-native speech corpus in a large scale. Moreover, it is much more difficult to precisely annotate non-native speech.

In this work, we propose methods to detect articulatory errors without using non-native training data. Effective articulation modeling of non-native speech is focused. Specifically, two large native speech corpora are used to model the cross-lingual acoustic features. For effective and efficient learning of DNN articulatory models, we also combine multi-task learning, which incorporates the phone-level information. Combination of the multi-lingual and the multi-task learning is realized in the proposed network architecture.

The rest of this paper is organized as follows: In Section 2, context-dependent articulation modeling with DNN is described. Section 3 presents a method to exploit two large native speech corpora to model the articulatory attributes. Section 4 addresses combining multi-task learning methods to enhance the DNN articulatory model. Section 5 and Section 6 respectively report the performance of these modeling and learning methods in the native attribute recognition task and the non-native pronunciation error detection task. Conclusions are in the final section.

2. CONTEXT-DEPENDENT ARTICULATION MODELING WITH DNN

Articulation means the movement of the tongue, lips, and other organs to make 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 investigate articulatory models to recognize the attributes of L2 learners. The L2 learners in this study are Japanese students who learn Mandarin Chinese. As a consequence, Mandarin and Japanese articulatory attributes are considered in this paper.

Articulatory attribute transcription is directly derived from the phone transcription. Considering the co-articulation effect, we aim to model all attributes in a context dependent way. As the mapping relation between attributes and phones is many-to-many, we prepare four kinds of articulatory transcriptions (manner, place-roundedness, place-backness and place-height) to represent all attributes. Fig. 1 gives an example of attribute labels mapped from phone labels. Similar to context-dependent tri-phones used in ASR, labels for tri-manners and tri-places are generated by taking into account the labels of neighboring attributes.

The DNN model uses 40-dimensional filterbanks plus their first and second derivatives. The input to the network is 11 frames, 5 frames on each side of the current frame. The DNN has 7 hidden layers with 2048 nodes per layer. DNN training consists of unsupervised pre-training and supervised fine-tuning.

Sentence	你好（HELLO）					
Phone	sil	n	i	h	ao	sil
Manner	sil	nasal	vowel	unvoiced-fricative	vowel	sil
Place & Backness	sil	alveolar	anterior	velar	central back	sil
Place & Height	sil	alveolar	high	velar	low middle	sil
Place & Roundedness	sil	alveolar	unroundedness	velar	unroundedness roundedness	sil

Fig. 1. Converting phone labels to articulatory labels.

3. ARTICULATORY ATTRIBUTE MODELING USING MULTI-LINGUAL LEARNING METHOD

Some of the articulation manners or places are shared among different languages, while others are different. For example, the place of phones /b, p, m/ is bilabial in both Chinese and Japanese. However, the stop consonants /p, t, k/ are pronounced with different manners of articulation. In Chinese, they are all aspirated stop while they are unvoiced stop in Japanese. As a result of the language transfer, it is hard for Japanese students to handle this new manner of articulation. They are prone to pronounce it without sufficient aspiration so that the phone sounds like its counterpart unaspirated one. Considering these, we try to model inter-language phenomena.

In this study, we adopt a multi-lingual DNN (ML-DNN) to exploit two large Chinese and Japanese native speech corpora to model the difference at the output layer while learning the feature extraction in the language-independent hidden layers. Fig. 2 shows how to train the bilingual manner using ML-DNN: two training samples (one is native Chinese /p/ with aspiration-manner, the other is native Japanese /p/ with unvoiced-manner) are sequentially presented to the network. Each frame is fed into the shared hidden layers and then the language-dependent output layer. Shared hidden layers can also be considered as an intelligent feature extraction module which aims at learning the bilingual feature representation. The configuration of ML-DNN is same as the standard DNN except for the language-specific output layers. This allows for learning non-native articulatory models without using a non-native speech data set.

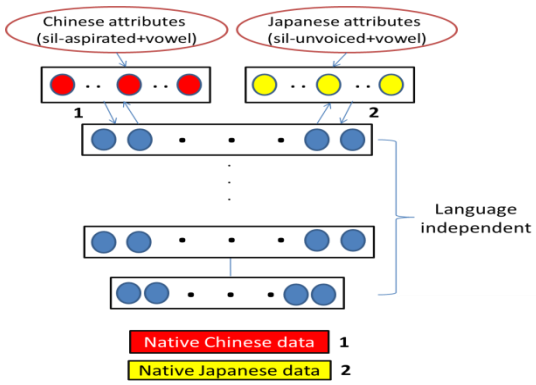

Fig. 2. Articulatory attribute modeling with ML-DNN.

4. ENHANCING ARTICULATORY ATTRIBUTE MODELING WITH MULTI-TASK LEARNING

4.1. Multi-task learning method

Multi-task learning is an approach that learns a problem together with other related problems at the same time. It has been successfully applied to various machine learning tasks [29-31]. In our study, different secondary tasks have been explored for enhancing the native articulatory attribute modeling. As a result, context-dependent tri-phone sub-task showed the best performance.

4.2. Combination of multi-lingual and multi-task learning methods

In this paper, we also investigate a combination of the above-mentioned multi-lingual learning and multi-task learning methods. There is not an established method for their combination and the simplest method is output-level combination such as recognizer output voting error reduction (ROVER) or lattice-level combination such as confusion network combination (CNC). In this paper, a new DNN architecture for network-level combination is designed

as shown in Fig. 3. It consists of shared hidden layers and language dependent output layers, which is similar to ML-DNN. However, the target language output layer is made of two tasks, i.e. phone classification and attribute classification tasks. This architecture allows the model to learn general features among different tasks and also different languages at the same time.

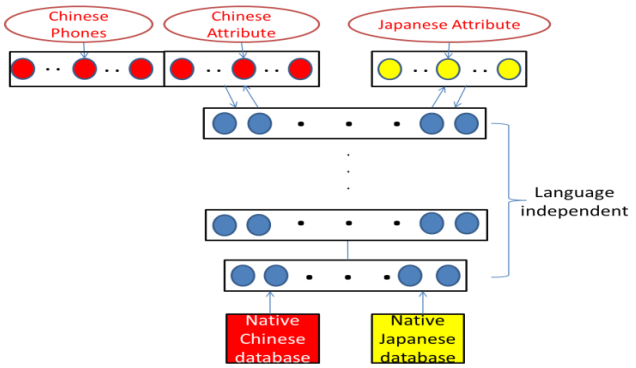

Fig. 3. Enhancing the articulatory models with phone classification sub-task.

5. NATIVE ATTRIBUTE RECOGNITION EXPERIMENT

The native Chinese and native Japanese speech corpora are used in this study. The Chinese corpus named db863, which is a corpus for speech recognition of Chinese National "863" Project. It has a total of about 110-hour recordings spoken by 166 speakers (83 females and 83 males). Mandarin Chinese is based on a particular Mandarin dialect spoken in the northern part of China, and almost same as the Beijing dialect. As our goal is to build a standard Chinese model, we use 64 speakers (36 females and 28 males) whose hometown is Beijing to train the standard articulatory model. We also use 8 speakers (5 males and 3 females) from the northern China for validating different methods. The duration for training and testing sets are about 42 hours and 5.3 hours. The Japanese corpus used for multi-lingual training is JNAS corpus. We also use a data set of 42 hours.

The recognition results of different DNN articulatory attributes are shown in Fig. 4. We observe the effect of multi-lingual DNN (ML-DNN) and multi-task DNN (MT-DNN) in the result, and MT-DNN is more effective than ML-DNN, because this evaluation is conducted on the Chinese native speech data only. However, we confirm the effect of combining ML-DNN (ML+MT), which exploits more data for general feature extraction.

6. PRONUNCIATION ERROR DETECTION OF L2 LEARNERS

6.1. Experimental setup

We apply the model to non-native speech for pronunciation

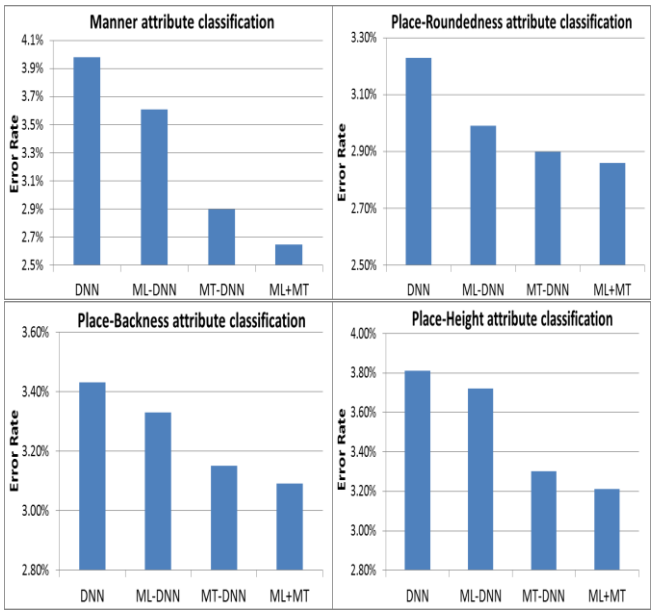

Fig. 4. Error rate of attribute classification.

error detection. The evaluation data used is continuous speech of the Japanese part in the BLCU inter-Chinese speech corpus, including 7 female speakers of Japanese native. All of them have learned Mandarin Chinese for many years and they all have an intermediate or advanced proficiency of Mandarin. Each learner uttered a same set of 301 daily-used sentences. There are 1896 utterances in total. The speech data were also annotated by 6 graduate students who majored in phonetics, and checked by a professor when they are inconsistent. The annotation contents are erroneous articulation described in [32]. For example, Chinese aspirated constant /p/ is pronounced with an incorrect articulation manner such as without meeting the required length of aspiration. Annotators will use a diacritic "p{;}" indicating this insufficient-aspiration error.

We employ finite state network decoding for pronunciation error detection, which includes the canonical pronunciation and possible pronunciation errors. Detection accuracy (DA) and F-score are used to evaluate the performance of different methods:

$$\text{DA} = \frac{N_{TE} + N_{TC}}{N}$$

$$\text{F-score} = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$\text{Precision} = \frac{N_{TE}}{N_D}$$

$$\text{Recall} = \frac{N_{TE}}{N_E}$$

N_{TE} is the number of true errors detected as pronunciation errors by the system. N_{TC} is the number of correct pronunciation detected as correct one by system. N is the total number of test samples. N_D is the number of all

detected pronunciation errors. N_E is the total number of pronunciation errors in the test set.

6.2. Experimental results

Fig. 5 compares the overall detection performance of five different methods: conventional DNN, MT-DNN, ML-DNN, and the combined ML+MT DNN (Section 4.2) and lattice-based combination of MT-DNN and ML-DNN. We can see that both MT-DNN and ML-DNN are better than the conventional DNN. While MT-DNN was consistently better than ML-DNN in the previous native attribute classification experiment, ML-DNN is generally more effective for modeling non-native speech. The combined ML+MT DNN further improved the performance. The combination of multi-lingual and multi-task learning on the network level shows better performance than the combination on the hypothesis level.

In this experiment, there are totally 4 pronunciation error types (involving 12 specific phones). All of them are typical and salient pronunciation errors even for advanced learners [33-35].

- Insufficient aspiration: insufficient aspiration when producing aspirated constants (e.g. p).
- Insufficient retroflex: insufficient retroflex when producing retroflex constants (e.g. zh).
- Lip rounding or spreading: vowels with spread lips have problems of rounded sound and vice versa (e.g. ü).
- Tongue backness: inappropriate tongue position with a little back (e.g. an).

Detailed detection results of individual error types are shown in Fig. 6. Among these 4 errors, the system detects the insufficient aspiration errors best, while the insufficient retroflex error detection is less accurate. This difference is mainly due to the subtle position difference among Mandarin retroflex, alveolar and palatal articulation placement. It should also be noted that pronunciation error detection of advanced learners is conducted in this study. Although with perceptual pronunciation errors, their articulation always deviates only a little from the canonical one. This brings a bigger challenge than detecting the errors made by beginners.

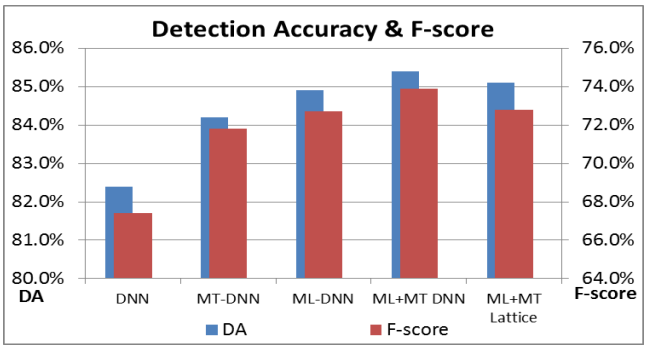

Fig. 5. Overall detection accuracy (DA) and F-score of different methods.

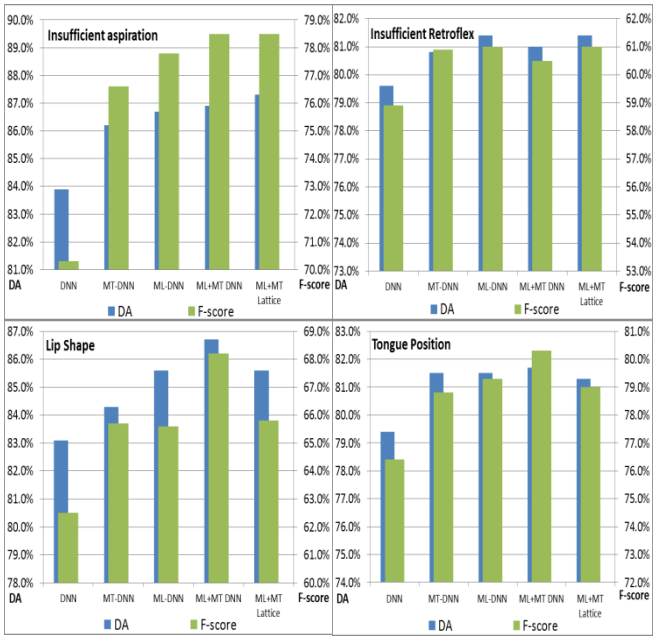

Fig. 6. Detection accuracy (DA) and F-score for 4 pronunciation error types.

7. CONCLUSIONS

For detecting articulatory pronunciation errors of second language learners' speech without using non-native training data, we propose methods that exploit two large native speech corpora to model articulatory attributes of non-native speech. Based on the shared articulatory attributes, the multi-lingual learning method is used to learn a better feature representation of non-native speech. A new model architecture is also introduced to improve generalization by allowing the model to jointly learn the general features among different tasks and different languages at the same time. Experimental results have shown that these approaches significantly improved the classification accuracy of native articulatory attributes and also detection of pronunciation errors produced by the learners.

In theory, the proposed approach can be applied to any language pairs as long as there is a native standard corpus. It opens new possibilities in language-independent pronunciation error detection.

REFERENCES

[1] C. Cucchiarini, F D. Wet, H. Strik, and L. Boves, "Assessment of Dutch pronunciation by means of automatic speech recognition technology," in Proc. ICSLP, 1998.

[2] C.-H. Jo, T. Kawahara, S. Doshita, and M. Dantsuji, "Automatic Pronunciation Error Detection and Guidance for Foreign Language Learning," in Proc. ICSLP, 1998.

[3] W. Menzel, D. Herron, P. Bonaventura, and R. Morton, "Automatic detection and correction of non-native English pronunciations," in Proceedings of Speech Technology in Language Learning, 2000.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

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

[11] 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.

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

[13] Y. Song, W. Liang, "Experimental Study of Discriminative Adaptive Training and MLLR for Automatic Pronunciation Evaluation," in Tsinghua Science & Technology, pp. 189-193, 2011.

[14] J. Zhang, F. PAN, B. Dong, Q. Zhao, and Y. Yan, "A Novel Discriminative Method for Pronunciation Quality Assessment," in IEICE, 96(5), pp. 1145-1151, 2013.

[15] 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.

[16] K. Truong, N. Ambra, C. Cucchiarini, and H. Strik, "Automatic pronunciation error detection: an acoustic-phonetic approach," in Proceedings of the 2004 InSTIL/ICALL Symposiumon on Computer Assisted Learning, pp.135-138, 2004.

[17] H. Strik, K. Truong, F. De Wet, and C. Cucchiarini, "Comparing classifiers for pronunciation error detection," in Proc. Interspeech, 2007.

[18] 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.

[19] Y. Tsubota, T. Kawahara, and M. Dantsuji. "Recognition and verification of English by Japanese students for computer-assisted language learning system," in Proc. ICSLP, 2002.

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

[21] Y.-B. Wang, L.-S. Lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training," in Proc. ICASSP, 2012.

[22] 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.

[23] A. Lee and J. Glass, "Context-dependent pronunciation error pattern discovery with limited annotation," in Proc. Interspeech, 2014.

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

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

[26] 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.

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

[28] 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.

[29] T. Cohn, L. Specia, "Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation," in Proc. ACL, 2013.

[30] W. Zhang, R. Li, T. Zeng, Q. Sun, S. Kumar, J. Ye, and S. Ji, "Deep model based transfer and multi-task learning for biological image analysis," in Proc. ACM, 2015.

[31] R. Rasipuram, M. Magimai-Doss, "Improving articulatory feature and phoneme recognition using multitask learning," in International Conference on Artificial Neural Networks, Springer Berlin Heidelberg, pp. 299-306, 2011.

[32] W. Cao, D. Wang, J. Zhang, and Z. Xiong, "Developing A Chinese L2 Speech Database of Japanese Learners With Narrow-Phonetic Labels For Computer Assisted Pronunciation Training," in Proc. Interspeech, 2010.

[33] X. Xie, "A study on Japanese Learner's Acquisition Process of Mandarin Balade-Palatal Initials," in Jilin Teachers Institute of Engineering and Technology, 2010.

[34] F. Li, W. Cao, "Comparative study on the acoustic characteristic of phoneme /u/ in mandarin between Chinese native speakers and Japanese learners," in Chinese Master's Thesis Full-text Database, No.S1, 2011.

[35] Y. Wang, X. Shanggguan, "How Japanese learners of Chinese process the aspirated and unaspirated consonants in standard Chinese," in Chinese Teaching in the World, 2004.