Multi-lingual and Multi-task DNN Learning for Articulatory Error Detection

Richeng Duan[*] Tatsuya Kawahara[*] Masatake Dantsuji[†] Jinsong Zhang[††]
[*] School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
[†] Academic Center for Computing and Media Studies, Kyoto University, Japan
[††] School of Information Science, Beijing Language and Culture University, China

Abstract—**For effective pronunciation error detection for second language learners, we address articulatory models based on deep neural network (DNN). Articulatory attributes are defined for manner and place of articulation. In order to efficiently train these models of non-native speech without using such data, which is difficult to collect in a large scale, we propose a multi-lingual learning method, in which the speech database of the target language (L2) and the native language (L1) of the learners are combined. We also investigate multi-task learning methods. These methods are applied to Mandarin Chinese pronunciation learning by Japanese native speakers. Effects of the multi-lingual and multi-task learning methods are demonstrated in the attribute classification of native speech and pronunciation error detection for non-native speech.**

I. Introduction

With the accelerating process of globalization, there is an increasing need for learning a second language (L2). It is every L2 learner's goal to have a correct and intelligible pronunciation. Computer-assisted pronunciation training (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 [1-6]. Students can study wherever and whenever they like. For effective learning, CAPT systems should give learners their pronunciation assessments and individualized corrective feedbacks.

In general, there are two main approaches to pronunciation assessment. One is to give learners pronunciation scores which involve from segmental level to speaker level, and the other detects individual errors such as specific phone substitution errors. The score in the sentence or speaker level can be measured over longer periods of time, and computed with a number of different phonetic and prosodic features. 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. For better pedagogical effects, the system should detect individual errors and provide corresponding feedbacks. This paper focuses on this problem, especially on the segmental aspects. Regarding the segmental pronunciation error detection, most of prior works focused on detection of phone substitution errors. Some researchers target a few specific problematic phones. They analyze the most frequent errors of those phones, and explore the distinctive features and classifiers. Others build systems with the automatic speech recognition (ASR) technology, either incorporating the possible errors into the lexicon or directly adding them into the decoding grammar. The ASR-based method is more general than the specially designed ones since it can detect any phones in a unified framework. However, it is not easy to reliably detect errors and to train the models of non-native speech. Moreover, detection of phone-level errors does not necessarily result in effective feedbacks for learners. In contrast, by using articulation information such as place and manner of articulation, we can provide feedbacks directly related with articulation, for example, "place your tongue in front" rather than giving "you mispronounced phone /l/ as /r/".

The above-mentioned detection methods need a non-native speech corpus to train statistical models, and the larger the corpus the better performance is expected. 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 a novel method to detect pronunciation errors without using non-native training data to provide feedbacks of the articulatory attributes. We achieve this primarily through modeling the place and the manner of articulation on the target language corpus. Context-dependent models of the articulatory attributes are defined using deep neural network (DNN). For effective and efficient learning of DNN articulatory models, we incorporate multi-task learning, which combines the phone-level classification task. Moreover, we also propose multi-lingual learning, in which the native language corpus of the learners is used since many articulatory attributes are shared between the two languages and we can easily get a large-scale native speech corpus. The effect of the model learning methods is evaluated in the articulatory attribute classification in the target and error detection in L2 learner's corpus.

The rest of this paper is organized as follows: we first describe models of the manner and place of articulation, then present multi-lingual and multi-task learning methods to enhance learning of the DNN articulatory models. The performance of these modeling and learning methods is first evaluated in native attributes recognition task. Then, it is applied to the non-native pronunciation error detection using the articulatory models.

II. 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 in the speech 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 used in this paper.

A. Articulatory attributes transcription

The place and manner transcriptions are derived from the phone transcription. Considering the many-to-many mapping relation between attributes and phones, we employ four kinds of articulatory transcriptions (manner, place-roundedness, place-backness and place-height) to represent all attributes. In the manner transcription, all vowels are mapped to the attribute named vowel. In other three place transcriptions, vowels are mapped into three-dimensional attributes respectively. Fig. 1 gives an example of attribute labels mapped from phone labels. Note that in Mandarin Chinese, there are compound vowels which are composed of more than one vowel. These compound vowels are mapped into the several attributes according to every single vowel. Hence the vowel “ao” in Fig. 1 is mapped into “unround round” attributes.

B. Context-dependent attribute modeling with DNN

We employ context-dependent tri-attribute modeling. 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 system 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	sil	你好（HELLO）				sil
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 Mid	sil
Place & Roundedness	sil	alveolar	unround	velar	unround round	sil

Fig. 1 Converting phone labels to manner and place-roundedness attribute labels.

III. MULTI-LINGUAL ARTICULATORY ATTRIBUTE MODELING

Different from the traditional DNN, there are more than one output layers in multi-lingual DNN (ML-DNN), and each language has its own output layer to compute the posterior probabilities. The hidden layers are shared by all languages and trained by all training samples, while each block output layer is only updated by language-dependent samples.

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 with a different manner of articulation. In Chinese, they are all aspirated stop while unvoiced stop in Japanese. As a result of the language transfer, when Japanese learners pronounce these aspirated stops, they may place it with a Japanese voiceless manner. Considering these, we adopt the ML-DNN learning to model the difference while learning the feature extraction in the language-independent hidden layers. The advantage of ML-DNN is to exploit two large corpora of native speech, Chinese and Japanese in this study, to model inter-language phenomena.

Fig. 2 shows how to train the bi-lingual manner using ML-DNN: Two training samples (one is Chinese /p/ with aspiration-manner, the other is Japanese /p/ with unvoiced-manner) are sequentially presented to the network. Each frame is fed into the shared hidden layers and the language-dependent output layer so that the hidden layers are trained for these two manners. Shared hidden layers can be considered as an intelligent feature extraction module which aims at learning the bilingual feature representation. The configuration of ML-DNN is same except for the language-specific output layers.

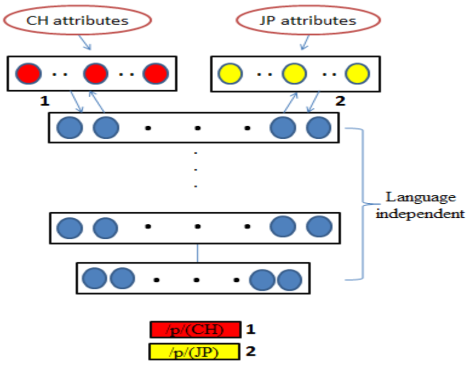

Fig. 2 Multi-lingual DNN for manner of articulation model

IV. MULTI-TASK LEARNING ON ARTICULATORY ATTRIBUTE MODELING

Multi-task learning is an approach of machine learning that learns a task together with other related tasks at the same time. Multi-task DNN (MT-DNN) has been successfully applied to various machine learning tasks. The aim of employing multi-task learning methods is effective and efficient learning of DNN articulatory models. The structure of MT-DNN is similar to ML-DNN, and they both have more than one output layers. However, all of the tasks are trained simultaneously in MT-DNN. In other words, both hidden layers and output layers are trained by all samples, which is different from the ML-DNN training process.

In current study, we use tri-phone classification as the secondary tasks for enhancing our primary task of articulation modeling.

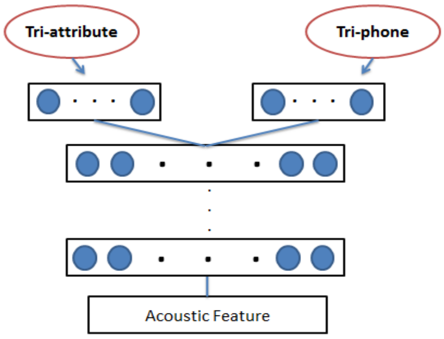

Fig. 3 Schematic diagram of multi-task DNN.

V. ATTRIBUTE RECOGNITION EXPERIMENT

The Chinese native speech corpus is primarily used for this study. Mandarin Chinese is based on particular Mandarin dialect spoken in the northern part of China, and almost same as Beijing dialect. As our aim is to build a standard Chinese model, we select 64 speakers (36 females and 28 males) whose hometown is Beijing to train the standard articulatory model. We also select 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, also about 42 hours.

The experimental results of different articulatory attributes are shown in Fig. 4. The weight of the secondary task is tuned with an interval of 0.2. The best result of different MT-DNN configuration is shown in the last bar of each attribute classification. From this figure, we see both ML-DNN and MT-DNN achieve better results than standard DNN. MT-DNN shows better performance than the ML-DNN, although it requires tuning of the weight.

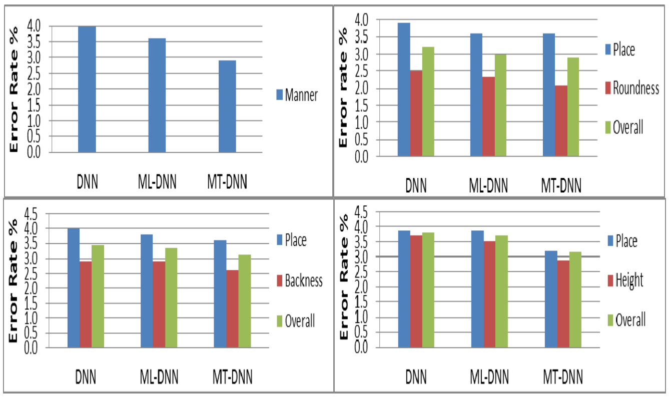

Fig. 4 Error rate of attributes classification.

VI. PRONUNCIATION ERROR DETECTION

A. *Experimental setup*

The evaluation data used in this work is continuous speech of the Japanese part in 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 recordings 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 tendencies described in [7]. For example, a 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{;}" this insufficient-aspiration error.

Here we use 2 metrics to evaluate the performance of error detection methods: Detection accuracy (DA) and F-score. DA is the proportion of true results (both true positives and true negatives) among the total number of cases detected. F-score is the harmonic mean of the precision and recall. Precision is the ratio of the number of true errors that are correctly identified by the classifier as pronunciation errors to the total number of hypothesized errors. Recall is the ratio of the number of true errors that are correctly detected to the total number of pronunciation errors in the test set

B. *Construction of detection graph*

In order to detect pronunciation errors, we employ a grammar-based graph for decoding, which includes the canonical pronunciation and possible pronunciation errors. Fig. 5 shows an example of how to construct a manner graph given the canonical pronunciation. The phone /t/ is an aspirated consonant in Chinese, while a voiceless constant in Japanese. As a result, it is hard for Japanese L2 learners to handle this new manner of articulation. Japanese learners are prone to pronounce it without sufficient aspiration so that the phone sounds like its counterpart unaspirated one. The aspirated manner and its counterpart can be represented in a finite state graph. We generate a detection graph for every sentence in this way.

Sentence:	你身体好吗？(How are you?)
Pinyin:	n i sh en t i h ao m a
Manner of Articulation:	nasal vowel unvoiced-fricative vowel aspirated-stop vowel unvoiced-fricative vowel nasal vowel

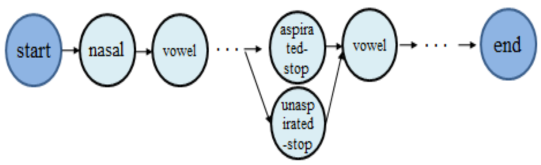

Fig. 5 Example of grammar-based detection graph.

C. *Experimental results*

In this paper, we focus on 4 pronunciation error tendencies which involve 12 specific phones:

- Insufficient aspiration: insufficient aspiration in manner of aspirated consonants.
- Insufficient retroflex: insufficient retroflex in place of articulation.
- Lip rounding or spreading: vowels with spread lips have problems of rounded sound and vice versa.
- Tongue Backness: inappropriate tongue position with a little back.

All of them are typical and salient pronunciation errors when Japanese speakers learn Chinese [8-10]. The overall detection performance of different methods is shown in Fig. 6. We can see that both ML-DNN and MT-DNN outperform the conventional DNN. While MT-DNN was consistently better than ML-DNN in the previous experiment using native speech only, ML-DNN is effective for modeling and classifying non-native speech. Fig. 7 gives detailed detection results of individual error types.

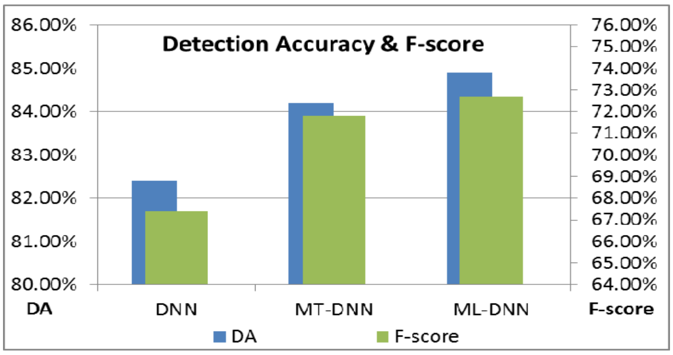

Fig. 6 Overall detection accuracy (DA) and F-score of different methods.

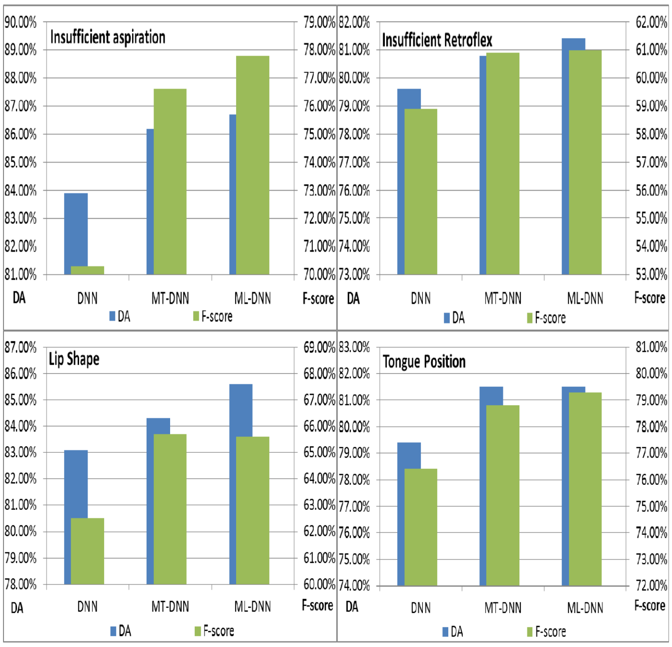

Fig. 7 Detection accuracy (DA) and F-score for 4 pronunciation error types.

VII. CONCLUSIONS

In this paper, we proposed employing multi-lingual and multi-task learning methods to model the articulation manner and articulation place for detecting the articulatory errors of L2 speech. The motivation of multi-lingual learning method is learning a better feature representation of L2 speech while multi-task learning is for effective and efficient learning of DNN articulatory models through inductive transfer. Experimental results have shown that these approaches significantly improved classification accuracy of articulatory attributes and also detection of pronunciation errors produced by L2 learners. These promising results lead us to investigate the combination of multi-lingual and multi-task learning methods in future work.

In theory, the proposed approach can be applied to any L1-L2 pairs as long as there is a native standard corpus. In future, we will try this approach on other language learning corpus, such as Chinese students learning English.

ACKNOWLEDGMENT

The author would like to thank for the financial support from Chinese Scholarship Council (CSC).

REFERENCES

[1] C.-H. Jo, H. Jo, T. Kawahara, S. Doshita, and M. Dantsuji, "Automatic Pronunciation Error Detection and Guidance for Foreign Language Learning," in ICSLP 1998, pp.2639--2642.

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

[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, 2008, pp. 160-167.

[4] T. Kawahara, H. Wang, Y. Tsubota, and M. Dantsuji, "English and Japanese CALL systems developed at Kyoto University," in APSIPA 2010, pp. 804-810.

[5] S. Joshi, N. Deo, P. Rao, "Vowel mispronunciation detection using DNN acoustic models with cross-lingual training," in INTERSPEECH 2015, pp. 697-701.

[6] 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.

[7] 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 INTERSPEECH 2010, pp. 1922-1925.

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

[9] 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.

[10] 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.