IEICE TRANS. INF. & SYST., VOL.E100–D, NO.9 SEPTEMBER 2017

PAPER

Articulatory Modeling for Pronunciation Error Detection without Non-Native Training Data Based on DNN Transfer Learning

Richeng DUAN[†a)], ***Nonmember***, **Tatsuya KAWAHARA**[†], ***Member***, **Masatake DANTSUJI**[††], ***and*** **Jinsong ZHANG**[†††], ***Nonmembers***

SUMMARY Aiming at detecting pronunciation errors produced by second language learners and providing corrective feedbacks related with articulation, we address effective 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 such data, which is difficult to collect in a large scale, several transfer learning based modeling methods are explored. We first investigate three closely-related secondary tasks which aim at effective learning of DNN articulatory models. We also propose to exploit large speech corpora of native and target language to model inter-language phenomena. This kind of transfer learning can provide a better feature representation of non-native speech. Related task transfer and language transfer learning are further combined on the network level. Compared with the conventional DNN which is used as the baseline, all proposed methods improved the performance. In the native attribute recognition task, the network-level combination method reduced the recognition error rate by more than 10% relative for all articulatory attributes. The method was also applied to pronunciation error detection in Mandarin Chinese pronunciation learning by Japanese native speakers, and achieved the relative improvement up to 17.0% for detection accuracy and up to 19.9% for F-score, which is also better than the lattice-based combination.

key words: *CALL, CAPT, pronunciation error detection, articulation modeling, transfer learning*

1. Introduction

With the accelerating process of globalization, there is an increasing need for learning a second language. Although one-to-one interactive lesson by experienced teachers is the most effective way, there is financial and time constraint for most students. With Computer-assisted Language Learning (CALL) systems, students can study wherever and whenever they like. Computer-assisted pronunciation training (CAPT) is an indispensable component of CALL system. For effective learning pronunciation, the system should provide learners their pronunciation assessments and individualized corrective feedbacks.

Over the last decades, CAPT based on statistical modeling techniques has made considerable progress [1]–[8]. There are two main approaches to pronunciation assessment. One is to give learners pronunciation scores which involve from segmental level to speaker level [9]–[15], and the other detects individual errors such as specific phone substitution errors [16]–[25]. The score in the sentence or speaker level can be measured over long 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. 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 [16]–[18]. 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 [19]–[25]. The ASR-based method is more general than the specially designed ones since it can detect any phones in a unified framework. 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 feedbacks, giving the feedbacks 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 language 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. We achieve this through modeling the place and the manner of articulation based on transfer learning. The idea of transfer learning, which should trace back to 20 years ago, has been successfully employed in broad research fields [29]–[32]. This study presents how to employ transfer learning on the articulation modeling of non-native speech with deep neural networks (DNNs). Inter-language transfer learning, related-task transfer learning, and combination of

Manuscript received January 12, 2017.
Manuscript revised April 13, 2017.
Manuscript publicized May 26, 2017.

[†]The authors are with the Graduate School of Informatics, Kyoto University, Kyoto-shi, 606–8501 Japan.
[††]The author is with the Academic Center for Computing and Media Studies, Kyoto University, Kyoto-shi, 606–8501 Japan.
[†††]The author is with the School of Information Science, Beijing Language and Culture University, Beijing, 100083 China.
a) E-mail: duan@sap.ist.i.kyoto-u.ac.jp

DOI: 10.1587/transinf.2017EDP7019

these two methods are explored. For effective and efficient learning of DNN articulatory models, three different related tasks are firstly investigated. In the language transfer learning method, two large native speech corpora of learners' native language (Japanese) and a target language (Chinese) are used to model the inter-language phenomenon since many articulatory attributes are shared between the two languages and we can easily get a large-scale corpus. In order to get benefit from both learning methods, their combination is realized in a new network architecture.

The rest of this paper is organized as follows: Context-dependent articulation modeling with DNN is firstly described in Sect. 2. Section 3 and Sect. 4 present the DNN articulatory modeling based on related task transfer and inter-language transfer learning. Section 5 addresses combining these two methods. Section 6 and Sect. 7 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 second language (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.

2.1 Articulatory Attribute Transcription

The place and manner transcription is derived from the phone transcription using mapping tables (Tables 1–3) which are made according to the rules [33], [34]. In these tables, Chinese attributes are presented first, followed by the shared attributes and Japanese attributes. Each consonant has one manner attribute and one place attribute, while vowels are described by the three dimensional attributes. Considering many-to-many mapping relation between attributes and phones, we model these attributes with four DNNs. In the manner DNN, all vowels are mapped to the attribute named vowel. In place DNN, vowels are mapped into three-dimensional attributes. Therefore, we build three place DNNs, i.e. place-backness DNN, place-height DNN, place-roundedness DNN. An example of attribute labels mapped from phone labels is shown in Table 4. Note that in Mandarin Chinese, there are compound vowels which are composed of more than one vowels. These compound vowels are mapped into the several attributes according to its

Table 1 Chinese (CH) and Japanese (JP) constant list with manner attributes.

Manner	Phone set	
Aspirated-stop	CH: p t k	/
Unaspirated-stop	CH: b d g	/
Aspirated-affricative	CH: c ch q	/
Unaspirated-affricative	CH: z zh j	/
Lateral	CH: l	/
nasal	CH: m n	JP: m n N
Voiced-fricative	CH: r	JP: w y
Unvoiced-fricative	CH: f s sh x h	JP: f s sh h
Unvoiced-stop	/	JP: p t k
Voiced-stop	/	JP: b d g
Unvoiced-affricative	/	JP: ts ch
Voiced-affricative	/	JP: z j
Flap	/	JP: r

Table 2 Chinese (CH) and Japanese (JP) constant list with place attributes.

Place	Phone set	
Retroflex	CH: zh ch sh r	/
Labiodental	CH: f	/
Bilabial	CH: b p m	JP: b p m
Alveolar	CH: d t n l z c s	JP: d t n r z ts s j ch sh
Palatal	CH: j q x	JP: y
Velar	CH: g k h	JP: g k N
glottal	/	JP: h

Table 3 Chinese (CH) and Japanese (JP) vowel list with place attributes.

Attribute	Place	Phone set	
Place-Backness	Anterior	CH: i ü	JP: i e
	Central	CH: a	JP: a u
	Back	CH: e u o	JP: o
Place-Height	High	CH: i u ü	JP: i u
	Mid	CH: o e	JP: e o
	Low	CH: a	JP: a
Place-Roundedness	Unroundedness	CH: a i e	JP: a i e
	Roundedness	CH: o u ü	JP: o

Table 4 Converting phone labels to articulatory labels.

Categoy	Transcription					
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 & Backness	sil	alveolar	high	velar	low middle	sil
Place & Backness	sil	alveolar	unroundedness	velar	unroundedness roundedness	sil

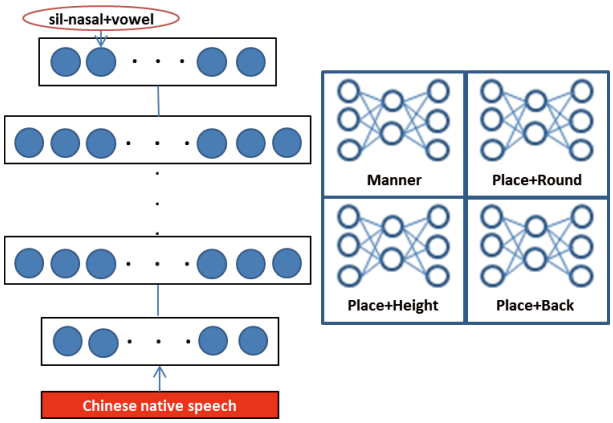

Fig. 1 Context-dependent modeling of articulatory attributes.

components. Hence the vowel "ao" in Table 4 is mapped into "unroundedness roundedness" attributes.

2.2 Context-Dependent Modeling Method for Articulatory Attributes

Considering co articulation effect, 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. We exploit Chinese native data to train the "standard" articulatory models (see Fig. 1). These target articulatory models can be directly used to detect pronunciation errors of L2 learners. This traditional DNN modeling method is served as our baseline method.

3. Multi-Task Learning on Articulatory Attribute Modeling

Multi-task learning is an approach of transfer 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 [35]–[37]. The goal is to improve the performance of learning algorithms by learning classifiers for multiple tasks jointly. It works particularly well if these tasks are closely related. Learning one task can help learning the other in the form of mutual regularization. Our aim of employing multi-task learning is effective and efficient learning of DNN articulatory models. The structure of MT-DNN is same to conventional DNN except for the multiple output layers.

In this study, we tried three different closely-related secondary tasks for enhancing our primary task. One is context-independent mono-attribute classification, which encourages the discrimination of the different attributes rather than different contexts of the same attribute. The other two are phone classification tasks (mono-phone and tri-phone classification), which aim at helping the primary task learn better feature representation of attributes with the phonetic information. We also investigate using different weights of the secondary task. Figure 2 gives the schematic diagram of MT-DNN training.

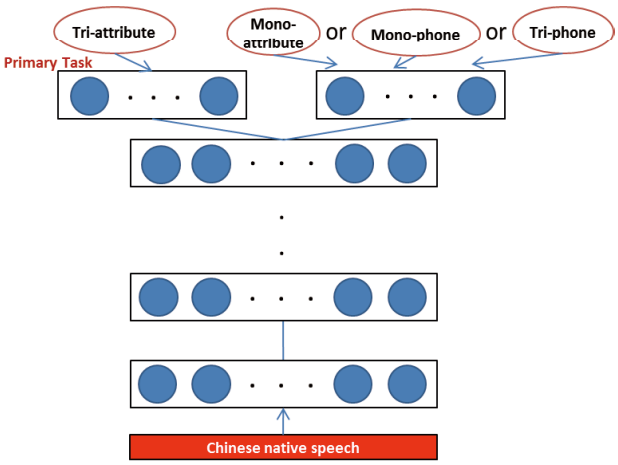

Fig. 2 Context-dependent modeling of articulatory attributes.

4. Multi-Lingual Learning on Articulatory Attribute Modeling

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. According to the language transfer theory [38]–[40], which refers to speakers applying knowledge from one language to another language, we know the following: When the relevant unit of both languages is the same, linguistic interference can result in positive language transfer. So Japanese students can easily pronounce an accented but correct place of Chinese phones /b, p, m/. On the other hand, when they are similar but not the same, negative transfer will occur. When Japanese students learning the Chinese aspirated consonants /p, t, k/, they are prone to pronounce them without sufficient aspiration and the phones sound like their native unvoiced ones.

Considering these, we investigate modeling of inter-language phenomena and learning it without non-native speech data. In the current 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 separated output layer while learning the commonality in the language-independent hidden layers. The structure of ML-DNN is similar to MT-DNN. However, all of the tasks are trained simultaneously in MT-DNN while the output layer is separately trained in ML-DNN. In other words, only hidden layers are trained by all samples in the ML-DNN training process. To be more specific, shared hidden layers can be considered as an intelligent feature extraction module which aims at learning the bilingual articulation representation. Features learned from this module coverage better acoustic characteristics of learners' speech. Figure 3 shows

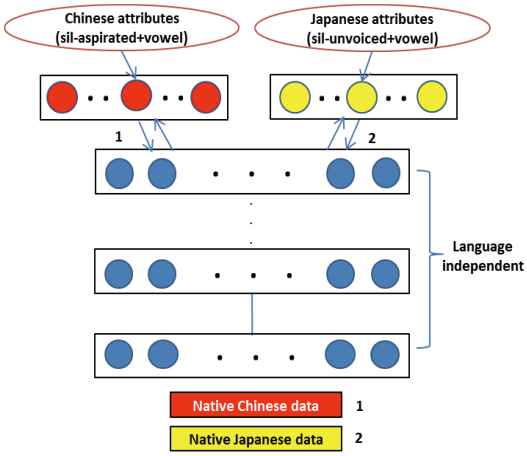

Fig. 3 Non-native articulatory attribute modeling with ML-DNN.

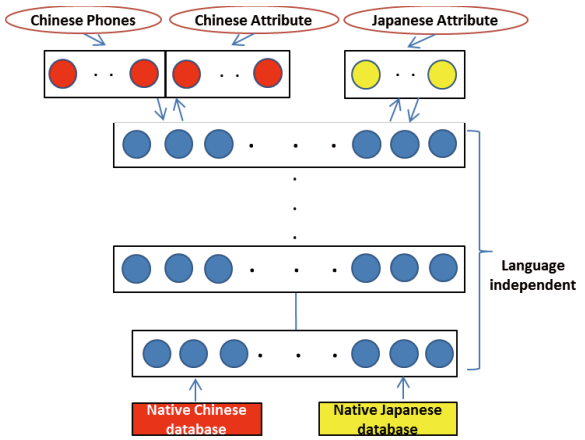

Fig. 4 Enhancing the articulatory models with network-level combination of MT-DNN and ML-DNN.

how to train the Chinese-Japanese 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 learn the commonality across these two languages while separated output layers learn the difference. This architecture allows for learning non-native articulatory features without using a non-native speech data set.

5. Enhancing Articulatory Attribute Modeling by Combining Multi-Lingual and Multi-Task Learning

Above mentioned multi-task and multi-lingual learning can be regarded as two particular implementation of transfer learning. Multi-task learning learns the commonality through co-supervision (each frame data has two labels). It can be seen as a kind of regularization approach on model level. Multi-lingual learning adopted here aims at learning a better feature representation of non-native speech. As a result, it is natural to investigate the combination of the two above-mentioned transfer learning methods.

The simplest method of combination is output-level combination such as recognizer output voting error reduction (ROVER) [41] or lattice-level combination such as confusion network combination (CNC) [42]. In this study, a new DNN architecture for network-level combination is designed as shown in Fig. 4. It also 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 kind of architecture allows the model to learn general features among different tasks and also different languages at the same time.

Table 5 Data set in native attribute recognition.

	Data set	Duration	#speakers
Train	Chinese native	42h	64
	Japanese native	42h	153
Test	Chinese native	5.3h	8

6. Native Attribute Recognition Experiment

6.1 Database

Two native speech corpora are used in this experiment. One is recorded by Chinese native speakers, which is used to train the standard articulatory models (DNN and MT-DNN) and validate different modeling methods. The other is recorded by Japanese native speakers. It is used in the ML-DNN and the combination model training.

The native Chinese corpus named db863, which is a corpus for speech recognition of Chinese National "863" Project [43]. 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 all the 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 evaluation. The duration for training and testing sets are about 42 hours and 5.3 hours. The Japanese corpus used for training is JNAS corpus [44], which is also a commonly used database for Japanese large-vocabulary continuous speech recognition research. We randomly select 42 hours speech data (80 males and 73 females) from it. All of these data sets are listed in Table 5.

6.2 System Configuration

All different methods use the following DNN configura-

tion: the acoustic feature consists of 40-dimensional Fourier transform based filter banks plus their 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. This configuration is determined by a preliminary experiment, in which we compared the performance of different numbers of layers (from 6 to 8) and nodes (1024 or 2048). DNN training consists of unsupervised pre-training and supervised fine-tuning.

6.3 Experimental Results

The experimental results of different articulatory attributes are shown in Fig. 5 to Fig. 8. From these 4 figures, we observe the effect of all three transfer learning based methods. Compared with the conventional DNN, all the methods achieve lower recognition error rates. Among the different configurations (different secondary task and weights) of multi-task learning method, the secondary task of context-dependent tri-phone improves the attribute recognition task most effectively. When we use the triphone as the secondary task, there is no significant difference in the performance among different weight values. We highlight the effect of combining multi-lingual and multi-task learning methods (ML+MT), which can reduce the recognition error rate by more than 10% relative in all articulatory attribute recognition tasks, though the conventional DNN has achieved a good performance with recognition error rate less than 5%.

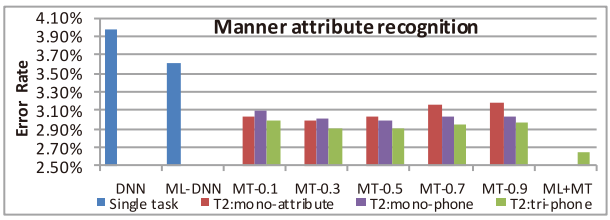

Fig. 5 Error rate of manner attribute recognition.

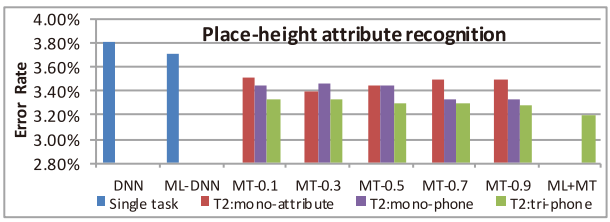

Fig. 6 Error rate of place-height attribute recognition.

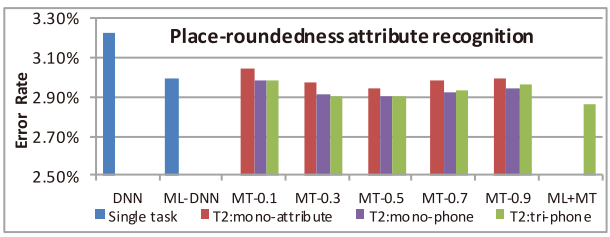

Fig. 7 Error rate of place-roundedness attribute recognition.

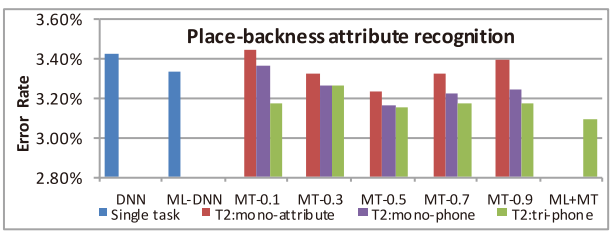

Fig. 8 Error rate of place-backness attribute recognition.

7. Pronunciation Error Detection of L2 Learners

7.1 Evaluation Database

The evaluation data for pronunciation error detection 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 [45]. For example, Chinese aspirated constant /p/ is pronounced with an incorrect articulation manner such as without meeting the required length of aspiration. Annotators used a diacritic "p{;}" indicating this insufficient-aspiration error. Table 6 gives some statistics of the database.

7.2 Construction of Detection Graph

We employ finite state network decoding for pronunciation error detection, which includes the canonical pronunciation and possible pronunciation errors. Figure 9 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. 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 as branching states in the decoding graph. We

Table 6 Summary of non-native evolution database.

Text	Conversational Chinese
Speaker	7 females
Number of utterance	1896
Number of phones	26534
Average length per utterance	14
Number of annotators	6

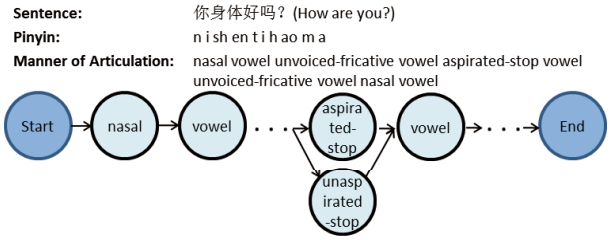

Fig. 9 Example of grammar-based detection graph.

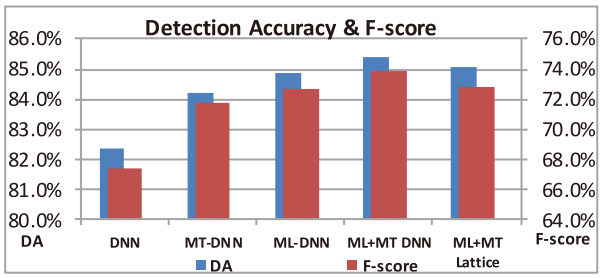

Fig. 10 Overall detection accuracy (DA) and F-score of different methods.

generate a detection graph for every sentence in this manner.

7.3 Evaluation Metrics

Two common used metrics Detection accuracy (DA) and F-score [15], [50] are used to evaluate the detection performance of different methods:

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

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

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

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

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 the 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.

7.4 Pronunciation Error Types

In this experiment, four pronunciation error types are focused which involve 12 specific phones:

- 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).

All of them are typical and salient pronunciation errors even for advanced learners [46]–[48]. By reviewing the sound systems shown in Tables 1–3, we can have an intuitive sense why these errors are representative. For example, comparing the manner of articulation in Chinese and Japanese (Table 1), we can see both Chinese and Japanese languages have the phones /p, t, k/. However, they are aspirated ones in Chinese while unvoiced in Japanese. As a result of language negative transfer, it will create a challenge for Japanese learners to mimic this new manner of articulation. Japanese learners are therefore prone to replace these aspirated phones with their native similar phones.

7.5 Experimental Results

Figure 10 compares the overall detection performance of five different methods: conventional DNN, MT-DNN, ML-DNN, the combined ML+MT DNN and lattice-based combination of MT-DNN and ML-DNN. In output level combination method, we assign a weight θ_i (from 0.1 to 1.0) to each system, where i = 1, 2 and $\theta_1 + \theta_2 = 1$. The combined lattice is generated by considering the Levenshtein edit distance [49]. We can see that both MT-DNN and ML-DNN are better than the conventional DNN, as observed in the native attribute recognition. MT-DNN improves DA by 10.2% relative and F-score 13.5% relative. ML-DNN improves the performance by 14.2% (DA) and 16.3% (F-score) relative. While MT-DNN is consistently better than ML-DNN in the previous native attribute classification experiment, ML-DNN is generally more effective for modeling non-native speech. This is because MT-DNN is trained with Chinese data only while we add Japanese characteristics by using both Chinese and Japanese data. The combined ML+MT DNN further improves the performance. The relative improvement is up to 17.0% for DA and up to 19.9% for the F-score. From this improvement, it is clear that MT-DNN and ML-DNN are complementary to each other though they are both transfer learning based methods. This network level combination of multi-lingual and multi-task learning shows better performance than the lattice combination on the output level (ML+MT Lattice).

Detailed detection results of individual error types are shown from Fig. 11 to Fig. 14. Among these four articulatory errors, the system detects the tongue backness error best shown in Fig. 13 (both DA and F-score are more than 80%), while the insufficient retroflex error detection is less accurate (F-score is about 60% shown in Fig. 12). Less accuracy in detecting the insufficient retroflex error is partly due to the subtle acoustic difference among Mandarin retroflex, alveolar and palatal articulation placement [50]. It should also be noted that pronunciation error detection of advanced learners is conducted in this study. Although with perceptual pronunciation errors judged by native speakers, their articulation deviates only a little from the canonical one. This brings a bigger challenge than detecting the errors made

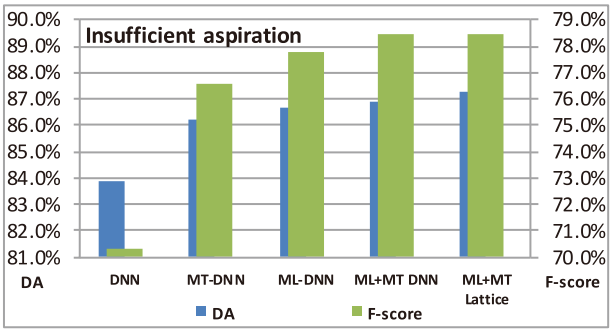

Fig. 11 Detection accuracy (DA) and F-score for insufficient aspiration error.

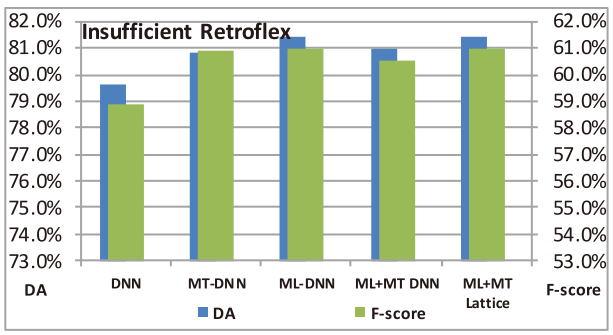

Fig. 12 Detection accuracy (DA) and F-score for insufficient retroflex error.

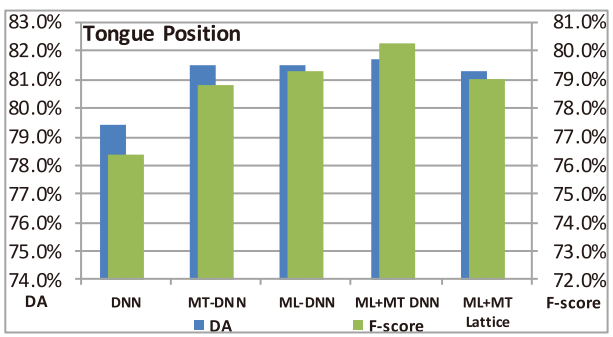

Fig. 13 Detection accuracy (DA) and F-score for tongue position error.

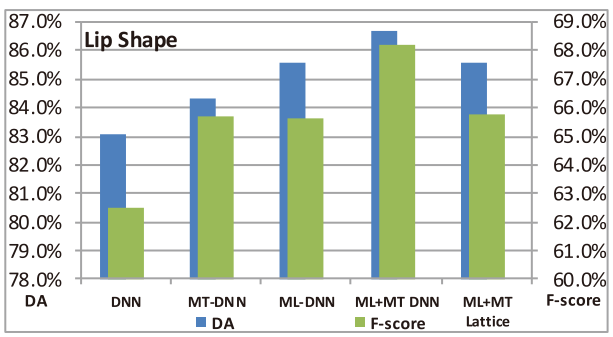

Fig. 14 Detection accuracy (DA) and F-score for lip shape error.

by beginners. A promising solution is designing and using more specific features, for example, adding the voice onset time (VOT) feature for discriminating stop consonants.

8. Conclusions

For detecting articulatory errors of second language learners' speech without using non-native training data, we propose to exploit large native speech corpora to model articulatory attributes of non-native speech. Several methods based on transfer learning are explored. The conventional DNN, which is used to model the standard articulatory attributes with Chinese native data, is firstly enhanced by multi-task learning. Multi-task learning improves the model training through co-supervision with different labels (attribute labels and phone labels). In order to include the Japanese learners' characteristic, another Japanese native corpus is added. Based on the articulatory attributes shared by these two languages, multi-lingual learning method is also explored, which aims at learning a better feature representation of non-native speech by language knowledge transfer (Chinese and Japanese). Finally, a new model architecture is introduced to make use of both transfer learning methods. This new architecture further improves generalization by allowing the model to jointly learn the commonality among different tasks and different languages at the same time. Experimental results have demonstrated that these approaches significantly improve 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. In future, we will apply these methods on Japanese or Chinese students learning English.

Acknowledgments

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

References

[1] C. Cucchiarini, F.D. Wet, H. Strik, and L. Boves, "Assessment of Dutch pronunciation by means of automatic speech recognition technology," Proc. ICSLP, pp.1739–1742, 1998.
[2] C.-H. Jo, T. Kawahara, S. Doshita, and M. Dantsuji, "Automatic Pronunciation Error Detection and Guidance for Foreign Language Learning," Proc. ICSLP, pp.2639–2642, 1998.
[3] W. Menzel, D. Herron, P. Bonaventura, and R. Morton, "Automatic detection and correction of non-native English pronunciations," Proc. Speech Technology in Language Learning, pp.49–56, 2000.
[4] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy technology interface in computer assisted pronunciation training," Proc. Computer assisted language learning, pp.441–467, 2002.
[5] S. Seneff, C. Wang, and J. Zhang, "Spoken conversational interaction for language learning," Proc. 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," Proc. 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," Proc. 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," Proc. IEEE/ACM Trans. Audio, Speech, Language Process., 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," Proc. Eurospeech, vol.2, pp.851–854, 1999.
[10] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," Proc. 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," Proc. ICASSP, p.IV-201, 2007.
[12] F. Zhang, C. Huang, F.K. Soong, M. Chu, and R.H. Wang, "Automatic mispronunciation detection for Mandarin," Proc. ICASSP, pp.5077–5080, 2008.
[13] Y. Song and W. Liang, "Experimental Study of Discriminative Adaptive Training and MLLR for Automatic Pronunciation Evaluation," Proc. 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," Proc. 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," Proc. 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," Proc. 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," Proc. Interspeech, pp.1837–1840, 2007.
[18] H. Strik, K. Truong, F. De Wet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," Proc. Speech Communication, vol.51, 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," Proc. ICSLP, pp.1205–1208, 2002.
[20] H. Meng, Y.Y. Lo, L. Wang, and W.Y. Lau, "Deriving salient learners' mispronunciations from cross-language phonological comparisons," Proc. ASRU, pp.437–442, 2007.
[21] Y.-B. Wang and L.-S. Lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training," Proc. ICASSP, pp.5049–5052, 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," Proc. ICASSP, pp.8232–8236, 2013.
[23] A. Lee and J. Glass, "Context-dependent pronunciation error pattern discovery with limited annotation," Proc. Interspeech, pp.2877–2881, 2014.
[24] A. Lee and J. Glass, "Mispronunciation Detection without Nonnative Training Data," Proc. Interspeech, pp.643–647, 2015.
[25] S. Joshi, N. Deo, and P. Rao, "Vowel mispronunciation detection using DNN acoustic models with cross-lingual training," Proc. Interspeech, pp.697–701, 2015.
[26] P. Badin, Y. Tarabalka, F. Flisei, and G. Baily, "Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding," Proc. 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," Proc. Interspeech, pp.2643–2646, 2008
[28] A. Rathinavelu, H. Thiagarajan, and A. Rajkumar, "Three dimensional articulator model for speech acquisition by children with hearing loss," Proc. 4th Internation Conference on Universal Access in Human Computer Interaction, vol.4554, pp.786–794, 2007.
[29] M.E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," Proc. Journal of Machine Learning Research, vol.10, pp.1633–1685, 2009.
[30] S.J. Pan and Q. Yang, "A survey on transfer learning," Proc. IEEE Trans. Knowl. Data Eng., vol.22, no.10, pp.1345–1359, 2010.
[31] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," Proc. ICML Unsupervised and Transfer Learning, pp.17–36, 2012.
[32] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, "Transfer learning using computational intelligence: A survey," Proc. Knowledge-Based Systems, vol.80, pp.14–23, 2015.
[33] J. Zhang, Chinese man-machine voice communication infrastructure, Shanghai Science and Technology Press, 2010.
[34] X. Pi, Japanese Summary, Shanghai Foreign Language Education Press, 1997.
[35] R. Rasipuram and M. Magimai-Doss, "Improving articulatory feature and phoneme recognition using multitask learning," Proc. International Conference on Artificial Neural Networks, pp.299–306, Springer Berlin Heidelberg, 2011.
[36] T. Cohn and L. Specia, "Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation," Proc. ACL, pp.32–42, 2013.
[37] 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," Proc. ACM SIGKDD, pp.1475–1484, 2015.
[38] L. Postman and K. Stark, "Role of response availability in transfer and interference," Proc. Journal of Experimental Psychology, 79.1p1, 168, 1969.
[39] J.D. Bransford, A.L. Brown, and R.R. Cocking, "How people learn: Brain, mind, experience, and school," Wasington, DC, National Academy Press, 1999.
[40] C.B. Chang and A. Mishler, "Evidence for language transfer leading to a perceptual advantage for non-native listeners," Proc. Journal of the Acoustical Society of America, vol.132, no.4, pp.2700–2710, 2012.
[41] J.G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp.347–354, 1997.
[42] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation, and system combination," Proc. Speech Transcription Workshop, vol.27, pp.78–81, 2000.
[43] S. Gao, et al., "Update of Progress of Sinphear: Advanced Mandarin LVCSR System At NLPR," Proc. ICSLP, pp.798–801, 2000.
[44] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," Proc. Journal of the Acoustical Society of Japan, vol.20, no.3, pp.199–206.
[45] 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," Proc. Interspeech, pp.1922–1925, 2010.
[46] X. Xie, "A study on Japanese Learner's Acquisition process of Mandarin Balade-Palatal Initials," Proc. Jilin Teachers Institute of Engineering and Technology, vol.26, no.7, pp.23–27, 2010.
[47] F. Li and W. Cao, "Comparative study on the acoustic characteristic of phoneme /u/ in mandarin between Chinese native speakers and Japanese learners," Proc. Chinese Master's Thesis Full-text Database, no.S1, pp.122–124, 2011.

[48] Y. Wang and X. Shangguan, "How Japanese learners of Chinese process the aspirated and unaspirated consonants in standard Chinese," Proc. Chinese Teaching in the World, vol.3, pp.54–56, 2004.
[49] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," Computer Speech & Language, vol.25, no.4, pp.802–828, 2011.
[50] 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," Proc. ICASSP, pp.6135–6139, 2016.

Richeng Duan received the B.S. degree in Computer Science, from Tianjin University of Technology, China, in 2011 and M.S. degrees in Information Science from Beijing Language and Culture University in 2015. His research interests include speech recognition, speech understanding and computer-assisted language learning. He is now a Ph.D. candidate in Kyoto University.

Tatsuya Kawahara received B.E. in 1987, M.E. in 1989, and Ph.D. in 1995, all in information science, from Kyoto University, Kyoto, Japan. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ, USA. Currently, he is a Professor in the School of Informatics, Kyoto University. He has also been an Invited Researcher at ATR and NICT. He has published more than 300 technical papers on speech recognition, spoken language processing, and spoken dialogue systems. He has been conducting several speech-related projects in Japan including speech recognition software Julius and the automatic transcription system for the Japanese Parliament (Diet). Dr. Kawahara received the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology (MEXT) in 2012. From 2003 to 2006, he was a member of IEEE SPS Speech Technical Committee. He was a General Chair of IEEE Automatic Speech Recognition and Understanding workshop (ASRU 2007). He also served as a Tutorial Chair of INTERSPEECH 2010 and a Local Arrangement Chair of ICASSP 2012. He is an editorial board member of Elsevier Journal of Computer Speech and Language, APSIPA Transactions on Signal and Information Processing, and IEEE/ACM Transactions on Audio, Speech, and Language Processing. He is VP-Publications (BoG member) of APSIPA and a Fellow of IEEE.

Masatake Dantsuji received the B.S. and M.S. degrees in Letters from Kyoto University in 1979 and 1981, respectively. During 1990-1997, he stayed in Kansai University as associate professor. From 1997, he is a professor of Kyoto University.

Jinsong Zhang received a B.E. in Electronic Engineering from Hefei University of Technology, China in 1989, an M. E. in Electronic Circuit, Signal and System from the University of Science and Technology of China (USTC) in 1992, and a Ph. D. in Information and Communication Engineering from the University of Tokyo, Japan in 2000. From 1992 to 1996 he worked as a teaching assistant and lecturer in the Department of Electronic Engineering at USTC. From 2000 to 2007, he had been an invited and senior researcher at ATR spoken language translation research laboratories. He is currently a professor in the school of computer sciences at Beijing Language and Culture University, Beijing, China. His research interests include speech recognition, prosody information processing, 2nd language acquisition, computer assisted pronunciation training, and etc.