Pronunciation Error Detection using DNN Articulatory Model based on Multi-lingual and Multi-task Learning

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
[3]School of Information Science, Beijing Language and Culture University, Beijing
duan@sap.ist.i.kyoto-u.ac.jp

Abstract

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 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 by tuning the weights of the secondary task. These methods are applied to Mandarin Chinese pronunciation learning by Japanese native speakers. Effects of the multi-lingual and multi-task learning methods are confirmed in the attribute classification and pronunciation error detection.

Index Terms: CAPT, pronunciation error detection, articulation modeling, multi-lingual learning, multi-task learning

1. 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 [1-8]. 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 [9-15], and the other detects individual errors such as specific phone substitution errors [16-24]. 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 [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 [19-24]. 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 a little back" rather than giving "you mispronounced phone /r/ as /l/".

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 the L2 learner's corpus.

The rest of this paper is organized as follows: In Section 2, models of the manner and place of articulation are described. Section 3 and Section 4 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 evaluated in Section 5. Section 6 reports the

pronunciation error detection using the articulatory models. Conclusions are in the final section.

2. 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.

2.1. Definition of articulatory attributes

The place and manner transcription is derived from the phone transcriptions using mapping tables (Table1-3). In these tables, Chinese attributes are presented first followed by the shared attributes and Japanese attributes. The attributes in boldface are shared by two languages. Each consonant has one manner attribute and one place attribute, while vowels are described by the three dimensional attributes. We model these attributes with independent deep neural networks (DNN). One DNN is for modeling the place attributes and the other is for the manner attributes. 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. Figure 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 Figure 1 is mapped into "unround round" attributes.

2.2. Context-dependent attribute modeling with DNN

We employ context-dependent tri-attribute modeling. Similar to context-dependent triphones 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 & Roundedness	sil	alveolar	unround	velar	unround round	sil

Figure 1: *Converting phone labels to manner and place-roundedness attribute labels.*

Table 1. *Chinese (CH) and Japanese (JP) vowel list with attributes*

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

Table 2. *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 3. *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

3. Multi-lingual articulatory attribute modeling

Different from the traditional DNN, there are more than one output layers in multi-lingual Deep Neural Network (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.

Figure 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 will be trained for these two manners. The configuration of ML-DNN is same except for the language-specific output layers.

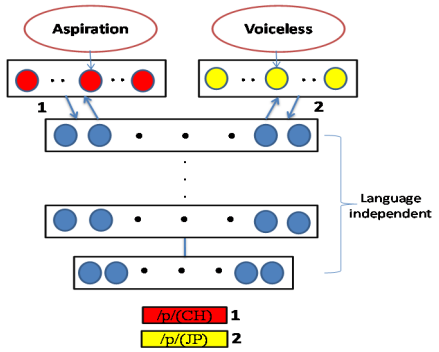

Figure 2: *Multi-lingual DNN for manner of articulation model*

4. Multi-task learning on articulatory attribute modeling

Multi-task learning [25] 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 [26-28]. 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.

Inspired by the previous work that integrated the articulatory knowledge into phone recognition using multi-task learning [29], we tried 2 different secondary tasks for enhancing our primary task of articulation modeling. One is context-independent mono-attribute classification, which is different from our primary task of context-dependent tri-attributes classification. The other is context-dependent phone classification. Moreover, we also investigate using different weights of the secondary task.

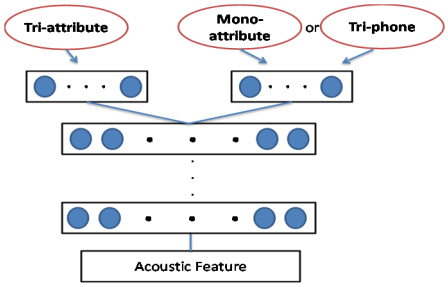

Figure 3: *Schematic diagram of multi-task DNN.*

5. 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 [30], also about 42 hours.

The experimental results of different articulatory attributes are shown in Figure 4 to Figure 7. We compared 4 different models (GMM, standard DNN, ML-DNN, and MT-DNN). From these 4 figures, we see MT-DNN achieves the best results in all attribute recognition tasks. Compared to mono-attribute, the secondary task of context-dependent triphones improved the primary task most effectively. The weight of the secondary task also influences the model performance, and all error rates have a tendency of decrease firstly then increase as we increase the weight with an interval of 0.2. We select the best one as our final MT-DNN model to detect the pronunciation errors in the next section.

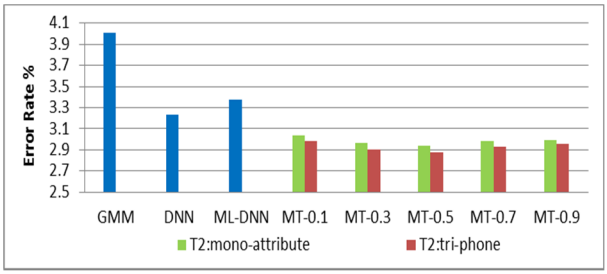

Figure 4: *Error rate of place-roundness attribute classification*

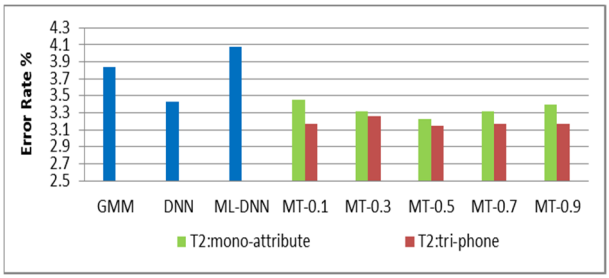

Figure 5: *Error rate of place-backness attribute classification*

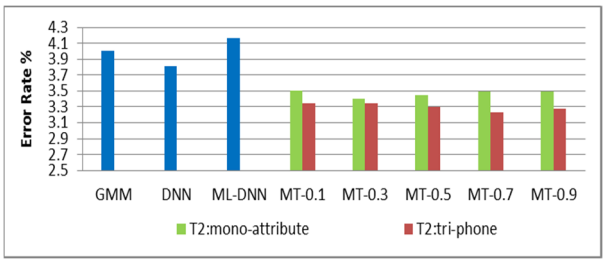

Figure 6: *Error rate of place-height attribute classification*

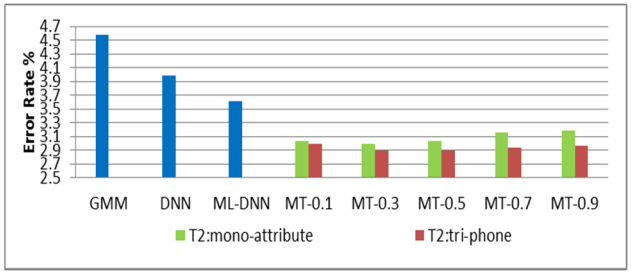

Figure 7: *Error rate of manner attribute classification*.

6. Pronunciation error detection

6.1. 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 professor when they are inconsistent. The annotation contents are erroneous articulation tendencies described in [31]. 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{;}" indicating this insufficient-aspiration error.

Here we use 3 metrics to evaluate the performance of error detection methods: False Alarm Rate (FAR), Miss Rate (MIS) and harmonic mean of these two error rates (HM).

6.2. 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. Figure 8 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

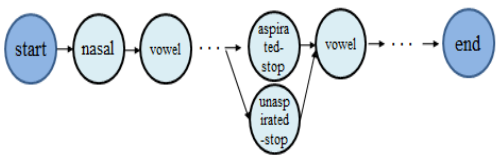

Figure 8: *Example of grammar-based detection graph.*

6.3. Experimental results

In this paper, we focus on 4 pronunciation error tendencies:

- Shortening: insufficient aspiration in manner of articulation.
- Laminalizing: insufficient retroflex in place of articulation.
- Lip rounding or spreading: sounds with spread lips have problems of rounded sound and vice versa.
- Backing: inappropriate tongue position with a little back.

All of them are typical and salient pronunciation errors when Japanese speakers learn Chinese [32-34]. The experimental results using different methods are shown in Table 4. In all blocks of Table 4, MT-DNN (multi-task DNN) achieves the best results, while ML-DNN (multi-lingual DNN) achieves improvements in two categories.

Table 4. *Detection error rate of different methods (%)*

	MODEL	FAR	MIS	HM
Shorting	GMM	29.9	9.4	14.3
	DNN	9.6	22.1	13.4
	ML-DNN	9.7	14.5	11.6
	MT-DNN	9.1	16.8	11.8
Laminalizing	GMM	5.2	53.3	9.5
	DNN	2.3	52.5	4.4
	ML-DNN	1.9	62.4	3.7
	MT-DNN	1.7	53.9	3.3
Lip rounding or spreading	GMM	19.7	28.7	23.4
	DNN	11.4	16.9	13.6
	ML-DNN	12.1	23.6	16.0
	MT-DNN	10.5	16.4	12.8
Backing	GMM	13.5	30.7	18.8
	DNN	9.9	29.5	14.8
	ML-DNN	12.0	30.5	17.2
	MT-DNN	10.3	25.5	14.6

7. 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. Experimental results have shown that this approach significantly improved classification accuracy of articulatory attributes and also detection of pronunciation errors produced by L2 learners.

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. We will also investigate more on the multi-lingual method for pronunciation error detection and combine it with the multi-task learning.

8. Acknowledgements

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

9. References

[1] C. Cucchiarini, F D. Wet, H. Strik, and L. Boves, "Assessment of dutch pronunciation by means of automatic speech recognition technology," in *ICSLP 1998, pp. 1739-1742*.

[2] C.-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*.

[3] W. Menzel, D. Herron, P. Bonaventura, and R. Morton, "Automatic detection and correction of non-native English pronunciations," in *Proc. of Speech Technology in Language Learning, 2000, pp. 49-56.*

[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 *Proc. of the InSTIL/ICALL Symposiumon on Computer Assisted Learning., 2004, pp. 151-154.*

[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, 2008, pp. 160-167*.

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

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

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

[10] S.Witt and S. Young, "Phone-level pronunciation scoring and assessment 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 *ICASSP 2007, pp. 201-204*.

[12] F. Zhang, C. Huang, F.K. Soong, M. Chu, and R.H. Wang, "Automatic mispronunciation detection for Mandarin," in *ICASSP 2008, pp. 5077–5080.*

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

[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 *Proc. of the 2004 InSTIL/ICALL Symposiumon on Computer Assisted Learning. 2004, pp.135-138.*

[17] H. Strik, K. Truong, F. De Wet, and C. Cucchiarini, "Comparing classifiers for pronunciation error detection," in *INTERSPEECH 2007, pp.1837-1840.*

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

[19] Y. Tsubota, T. Kawahara, and M. Dantsuji. "Recognition and verification of English by Japanese students for computer-assisted language learning system," in *ICSLP 2002, pp.1205--1208*.

[20] H. Meng, Y. Lo, L. Wang, and W. Lau, "Deriving salient learners' mispronunciations from cross-language phonological comparisons," in *Proc. ASRU 2007, pp.437-442.*

[21] Y.-B. Wang, L.-S. Lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training," in *ICASSP 2012, pp. 5049–5052*.

[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 *ICASSP 2013, pp. 8232-8236.*

[23] A. Lee and J. Glass, "Context-dependent pronunciation error pattern discovery with limited annotation," in *INTERSPEECH 2014, pp. 2877-2881.*

[24] A. Lee and J. Glass, "Mispronunciation Detection without Nonnative Training Data," in *INTERSPEECH 2015, pp. 643-647.*

[25] R. Caruna, "Multitask learning: A knowledge-based source of inductive bias," in *Machine Learning: Proceedings of the Tenth International Conference, 1993, pp. 41–48*.

[26] T. Cohn, L. Specia, "Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation," in *ACL 2013, pp. 32-42*.

[27] 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 *ACM 2015, pp. 1475–1484*

[28] D. Chen, B. Mak, CC. Leung, and S. Sivadas, "Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition," in *ICASSP 2014, pp. 5592–5596.*

[29] H. Zheng, Z. Yang, L. Qiao, J. Li, and W. Liu, "Attribute Knowledge Integration for Speech Recognition Based on Multi-task Learning Neural Networks," in *INTERSPEECH 2015, pp.543-547*

[30] 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," in *Journal of the Acoustical Society of Japan, Vol 20, pp. 199-206, 1999.*

[31] 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*.

[32] 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*.

[33] 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*.

[34] 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*.