Cross-Lingual Transfer Learning of Non-Native Acoustic Modeling for Pronunciation Error Detection and Diagnosis

AUTHOR(S):

Duan, Richeng; Kawahara, Tatsuya; Dantsuji, Masatake; Nanjo, Hiroaki

Cross-Lingual Transfer Learning of Non-Native Acoustic Modeling for Pronunciation Error Detection and Diagnosis

Richeng Duan, *Non-Member, IEEE*, Tatsuya Kawahara, *Fellow, IEEE*,
Masatake Dantsuji, *Non-Member, IEEE*, Hiroaki Nanjo, *Member, IEEE*

***Abstract*—In computer-assisted pronunciation training (CAPT), the scarcity of large-scale non-native corpora and human expert annotations are two fundamental challenges to non-native acoustic modeling. Most existing approaches of acoustic modeling in CAPT are based on non-native corpora while there are so many living languages in the world. It is impractical to collect and annotate every non-native speech corpus considering different language pairs. In this work, we address non-native acoustic modeling (both on phonetic and articulatory level) based on transfer learning. In order to effectively train acoustic models of non-native speech without using such data, we propose to exploit two large native speech corpora of learner's native language (L1) and target language (L2) to model cross-lingual phenomena. This kind of transfer learning can provide a better feature representation of non-native speech. Experimental evaluations are carried out for Japanese speakers learning English. We first demonstrate the proposed acoustic-phone model achieves a lower word error rate in non-native speech recognition. It also improves the pronunciation error detection based on goodness of pronunciation (GOP) score. For diagnosis of pronunciation errors, the proposed acoustic-articulatory modeling method is effective for providing detailed feedback at the articulation level.**

***Index Terms*—CALL, CAPT, non-native acoustic modeling, pronunciation error detection and diagnosis, cross-lingual transfer**

I. Introduction

COMPUTER-ASSISTED language learning (CALL) system is becoming more and more popular due to its flexibility of allowing students to practise their language skills in a stress-free environment at their convenient time and pace. CAPT as an indispensable component of the CALL system aims to improve learner's speaking skill through providing corrective feedbacks for pronunciation errors just like an experienced teacher does. To provide useful pronunciation instructions, the CAPT module needs to perform pronunciation error detection and diagnosis. Pronunciation errors are usually characterized at the phonetic (segmental) and prosodic (suprasegmental) levels. We focus on phonetic pronunciation errors in this paper. In terms of the diagnosis for detected errors, one main approach is identifying incorrect phones produced by the learner [1]. A typical feedback based on this approach is "You made an r-l substitution error" when a student pronounces the word "red" as "led". Instead of specifying the phoneme uttered in place of the canonical one, exploiting information directly related with articulation is more attractive because it provides corrective feedback of how to move the articulators in order to produce the target sound. Facing the same pronunciation error described above, learners would be instructed with "Try to retract your tongue when speaking the 'r' sound". Various kinds of articulatory attributes have been explored by the researchers in CAPT research field [2]-[5]. And this approach has been demonstrated more helpful in many areas, such as pronunciation perceptual training [6], speech therapy [7], and speech comprehension improvement [8]. Conducting diagnosis on the articulation level is focused in our work as a result.

As pronunciation of a foreign language is easily affected by the learners' native language, it is better to train acoustic model with the learners' speech data of target foreign language. However, it is much more difficult to collect and label non-native speech than native speech because of the fewer user populations and unnatural pronunciations [9]. While automatic speech recognition (ASR) has recently achieved great progress due to the emergence of Deep Neural network (DNN) and big data, DNN-based CAPT cannot benefit a lot because of the scarce of a large amount of training data. To overcome the problem of lacking large-scale annotated resources, we have explored several transfer learning based methods for Mandarin Chinese learning in [10] and several knowledge combination strategies in [11] which aim at effective learning of DNN articulatory models of non-native speakers without using such training data. The proposed cross-lingual transfer-learning is essentially multi-lingual training of DNN using the target

R. Duan and T. Kawahara are with the Graduate School of Informatics, Kyoto University, Kyoto-shi, 606–8501 Japan. (e-mail: duanricheng@gmail.com; kawahara@i.kyoto-u.ac.jp).

M. Dantsuji and H. Nanjo are with Academic Center for Computing and Media Studies, Kyoto University, Japan (e-mail: mdantsuji@media.kyoto-u.ac.jp; nanjo@media.kyoto-u.ac.jp).

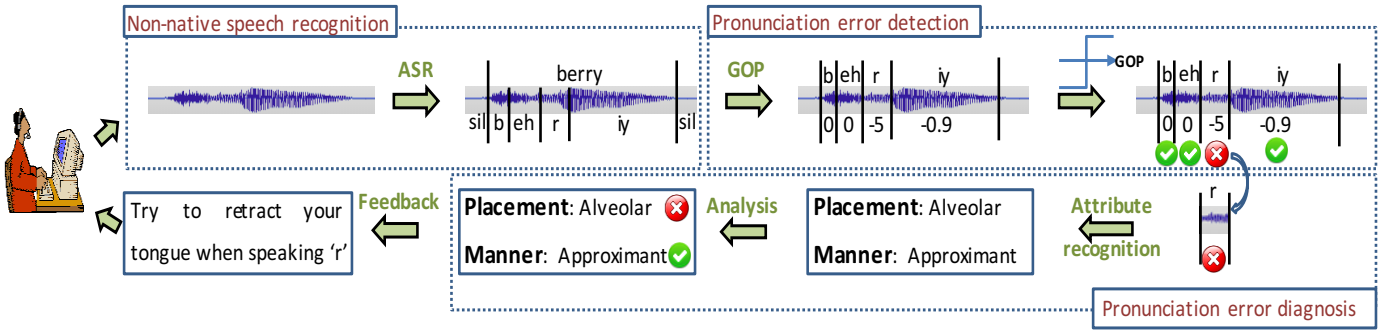


Fig. 1. Framework for pronunciation error detection and diagnosis. Learner's speech is first sent to the ASR engine which outputs the recognized text. For example, a single word "berry". Then its canonical phone sequence (b eh r iy) are used to calculate the recognition confidence score, such as the GOP score which is a variation of phone posterior probability. After that, the system can specify which phone is an error using a predefined threshold. The detected phone segment error is then sent to the diagnosis part which is a bank of articulatory attribute recognizers. The recognized attributes (Alveolar and Approximant) will be compared against the canonical attributes (Palato-alveolar and Approximant) of the target phone, and the difference (Alveolar vs. Palato-alveolar) reflects the cause of pronunciation error. The feedback related to articulators is naturally formulated as shown in this figure.

language dataset and the learners' native language dataset. Cross-lingual knowledge transfer was shown the most effective among those different kinds of prepared knowledge. Notice that this is different from the conventional cross-lingual or multi-lingual transfer applied to ASR [12]-[15] in that we do not assume any dataset of the target non-native speech that will be used during testing, but only assume the native speech datasets. In this paper, we extend proposed cross-lingual transfer learning [10] to a new L2 language (English), aiming to investigate its generality. In addition, the transfer learning is applied to not only the articulatory level [10] but also the phonetic level. Their effects will be examined in a newly designed CAPT system, in which the acoustic-phonetic model is used to recognize the non-native speech and perform pronunciation error detection while the acoustic-articulatory model is to conduct diagnosis of pronunciation errors. This framework allows for detection and diagnosis of any error patterns while our previous work [10] [11] based on a single-pass framework could be applied to only pre-defined limited patterns. We will present how the phone model and articulatory models can be effectively trained based on the cross-lingual and multi-task transfer-learning, and evaluate their effect on three steps of non-native speech recognition, pronunciation error detection, and articulatory error diagnosis.

The rest of this paper is organized as follows: In Section II, we review previous related work on pronunciation error detection and diagnosis. Section III introduces the acoustic modeling with conventional DNN based method. We present our proposed cross-lingual transfer learning based acoustic modeling approach in Section IV. Section V describes the speech corpora used for acoustic model training and evaluation. Section VI reports experiment evaluations in native attribute recognition and three non-native speaker related tasks of non-native speech recognition, pronunciation error detection and diagnosis. Conclusions are in the final section.

II. Related Work

While a limited number of studies have been conducted on unconstrained spontaneous speech [16] [17], most of previous works in CAPT are based on read speech where it assumes text-dependence [18]-[22]. The major challenge to achieve a text-independent system comes from the difficulty of non-native speech recognition. However, as the students improve, especially for those advanced learners, it would be better to let them speak freely and create their own sentences as opposed to reading a given text. We focus on text-independent CAPT and show the framework of proposed system in Fig. 1. The system includes four parts: First is recognizing the learner's speech; the next two are detecting the pronunciation errors using the recognized text and diagnosing the causes for errors. The last part is to provide articulatory feedbacks based on the diagnostic result.

A. Non-native Speech Recognition

To support text-independent CAPT, the system needs to recognize non-native speech in the first place irrespective of any pronunciation errors. However, recognition accuracy has been observed to be drastically lower for non-native speakers than for the native speakers [23]-[26]. This is mainly because the non-native speakers' pronunciation often differs from the native speech used in acoustic model training. Pronunciation errors, non-native accents, and disfluent speech all pose substantial difficulties for ASR. The most straightforward approach to tackle the mismatch problem is to use non-native speech data to train the acoustic model [25]. However, this kind of data is scarcely available and expensive to collect. An alternative approach has been proposed for acoustic model adaptation with limited non-native training data. The adaptation in [26] is conducted by freezing the lower layers of DNN while only the output layer is updated with a non-native dataset. Another popular approach to cover pronunciation variations is to construct a non-native pronunciation lexicon in

which the pronunciation of each lexical item is augmented with multiple pronunciations. It is constructed by using either linguistic rules which are derived through analyzing phonological structures of each L1-L2 pair [27], or a data-driven approach [28].

B. Pronunciation Error Detection

Pronunciation error detection is used to specify the correctness for each target phoneme. A pre-defined script or the recognized text (in a text-independent scenario) is used to perform forced alignment and to calculate the pronunciation 'correctness'. The most widely-used approach is based on ASR confidence measures which show the probability of correctness per speech segment. Up to now, various types of confidence measures, such as likelihood score [29], posterior probability [30], and likelihood ratio [31] have been investigated for pronunciation error detection. With the improvement in acoustic modeling, GOP (goodness of pronunciation) score, a variation of posterior probability, has been extensively adopted from traditional GMM-based acoustic model [32] to the current DNN-based system [33]. The decision to specify a pronunciation error is made by thresholding the GOP score where the thresholds are determined empirically. Considering native acoustic model is not well suitable to non-native testing samples, KL adaptation techniques were explored to reduce the mismatch [34]. In this study, we adopt the DNN-based GOP method as it can be obtained easily with the ASR system and assembled into our system quickly.

C. Pronunciation Error Diagnosis

From a pedagogical point of view, the CAPT system should be capable of not only pinpointing pronunciation errors, but also diagnosing the causes in order to provide corrective feedback to learners. Some prior studies focus on specific phoneme pair errors that are frequently made by foreigners, and design a corresponding classifier to identify errors spoken by the learner [35]-[37]. Instead of defining a special set of classification targets, a more general approach is directly recognizing the surface phoneme sequence produced by language learners. This implies we need to recognize erroneous non-native speech while simultaneously detecting errors. One of the most popular approach is one-pass pronunciation error detection and diagnosis with an "extended recognition network" (ERN) [38]. The ERN is usually constructed with a customized pronunciation lexicon. In addition to the canonical pronunciation per word, it explicitly incorporates all possible phonetic error patterns into the lexicon. The incorporated error patterns can be found from the knowledge of each L1-L2 pair. Some studies consult experienced expertise or carry out phonological comparisons between the L1-L2 pair [39]-[41] while others adopt data-driven approaches [42] [43]. Though above approaches are able to provide corrective feedback, the performance heavily relies on the quality of constructed error patterns. Moreover, one-pass based approach cannot easily select optimal models and take the "cost/benefit" context into consideration. On the other hand, the two-pass framework [1] detects the places where there are possible errors in first pass. In the second pass, phone loop recognition is conducted at the problematic places to identify the actual error types. We conduct articulatory-level diagnosis through articulatory attributes loop recognition in the second pass. The main advantage of this framework is that the detection pass can be used to control the system risk naturally (e.g. false alarm rate should be lower in the context of CAPT) through varying the threshold value.

III. Acoustic Modeling With Conventional DNN

As introduced in Section II, acoustic model is an essential component employed in CAPT systems. In this paper both acoustic-phonetic and acoustic-articulatory models are designed and implemented. The acoustic-phonetic model is used in non-native speech recognition and pronunciation error detection while the acoustic-articulatory model is to conduct diagnosis on learner's articulation. Articulation means the movement of the tongue, lips, and other organs to make speech sounds.

A. Phonetic and Articulatory Attributes Transcription

The phone level transcription is derived from the word sequence using the CMU pronunciation dictionary[1]. For the articulatory transcription, various methods have been investigated to generate speaker's articulatory attributes, including X-rays [44], electromagnetic articulography (EMA) [45], magnetic resonance imaging (MRI) [46], and ultrasounds [47]. However, all of the above direct measurements have disadvantages [48]. Above all, it is not easy to obtain such kinds of physical resources, especially in a large scale. In present work, we derive the attribute transcriptions from the phone transcription according to the phone-to-attribute mapping rules, which is a practical option adopted by many researchers [49]-[53]. 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). From the example in Fig. 2, we can see the mapping relation between the phone class and the attribute class is many-to-many (phone /M/ has two attributes nasal and bilabial while both vowels /IH/ and /AX/ are mapped to the unrounded attribute). Therefore, we

Word	sil	MR.						sil
English phone	sil	m	ih	s	t	ax	r	sil
Manner	sil	nasal	vowel	unvoiced-fricative	unvoiced-stop	vowel	approximant	sil
Place & Backness	sil	bilabial	anterior	alveolar	alveolar	central	palato-alveolar	sil
Place & Height	sil	bilabial	mid	alveolar	alveolar	mid	palato-alveolar	sil
Place & Roundedness	sil	bilabial	unround	alveolar	alveolar	unround	palato-alveolar	sil

Fig. 2. Converting word labels to phone labels and articulatory attribute labels.

[1] http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/

prepare four kinds of transcriptions (manner, place-roundedness, place-backness and place-height) to represent all articulatory attributes. In each kind of transcription, the attributes are disjoint to each other so that it can be used to train a DNN model.

B. DNN based Acoustic Modeling using L2 Native Data

Inspired by the great success of DNN based acoustic modeling in ASR, we follow the conventional DNN [54] to train both phonetic and articulatory models using the L2 native data (see Fig. 3). The language learners in this study are Japanese students who learn English. As a consequence, the acoustic models are trained from English native speech database. Considering the co-articulation effect, we model all acoustic units (phone and attribute) in a context dependent way. Tri-phone is used to train the acoustic-phonetic model while tri-attribute unit (e.g. tri-manner) is adopted for the acoustic-articulatory model. All of them are generated by taking into account the labels of neighboring phones or attributes. The targets in the output layer are the senone states and obtained by using a baseline GMM-HMM system to produce a forced alignment. A bank of DNNs are usually adopted [49] [51] because of the many-to-many mapping relation discussed above. In this work, we trained four

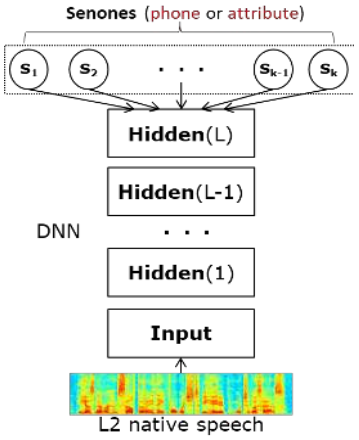


Fig. 3. Diagram of DNN based acoustic modeling for phones and attributes.

articulatory DNNs in which each DNN was used to represent one-kind attribute shown in Fig. 2. The articulatory models are used for diagnosis of learners' pronunciation, while the phone model is used for non-native speech recognition and pronunciation error detection in our CAPT system depicted in Fig. 1.

IV. Acoustic Modeling Based On Transfer Learning

The idea of transfer learning (TL), which traces back to 20 years ago, has been successfully employed in broad research fields [55]-[60]. Two major issues in TL are what knowledge to transfer and how to transfer. In order to enhance the acoustic model of non-native learners, two kinds of knowledge are investigated and compared in this paper. In terms of how to transfer, we assume the DNN consists of shared hidden layers

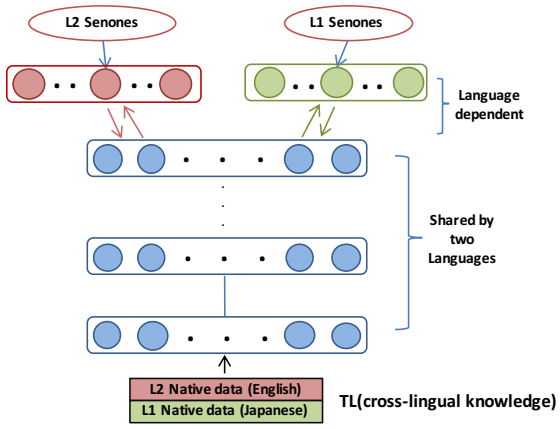


Fig. 4. Diagram of transfer learning based acoustic modeling using cross-lingual knowledge. Acoustic characteristic of language learners is learned through transferring knowledge of the learners' native language. Senones are prepared for both phones and articulatory attributes

and task-dependent output layers in this study, and shared hidden layers are used to transfer the knowledge of the source task to our target task.

A. Cross-Lingual Transfer using L1 and L2 Native Corpora

According to the language transfer theory [61]–[63], which refers to speakers applying knowledge from one language to another language, we assume the following: when the relevant aspect of both languages is same or very similar, linguistic interference can result in positive language transfer. On the other hand, when they are only comparable but not similar enough, or the linguistic unit in L2 is absent from L1, negative transfer will occur. For Japanese students learning English, they can easily pronounce an accented but correct English consonants /p, k, s, z/, which are shared by the two languages, while pronunciation becomes much more difficult when they turn to the English vowels, most of which are not present in Japanese. When we compare the vowel inventory, there are only five vowels in Japanese language while sixteen vowels (including the schwa sound) are in English. This significant phonological difference between the Japanese vowel system and the English one pose a big challenge for Japanese students.

Based on the language transfer theory, we propose to model the cross-lingual phenomena by exploiting two large native speech corpora (English and Japanese in this study) and employing the DNN structure made of shared hidden layers and separate output layers. The positive transfer is expected to be learned through shared hidden layers, while we model the differences at the separated outputs. Fig. 4 shows the designed DNN structure. Different from the traditional modeling method in Section III which only use L2 native dataset, both L1 and L2 native data are used during the training process. Each frame is firstly fed into the shared hidden layers and then its corresponding language-dependent output layer. During backpropagation, hidden neurons are then trained by two language samples while the gradient values of neurons in the output layer are fixed to zero if the language ID of current input is different from the output layer language ID. Assuming that

there are N speech samples in a minibatch, the loss function is defined as:

$$loss_{TL(cross-lingual)} = \frac{1}{N}\sum_{i=1}^{N}\left(I_{LID}(i)loss_{L2}^{i} + \left(1 - I_{LID}(i)\right)loss_{L1}^{i}\right) \quad (1)$$

where $loss_{L2}$ and $loss_{L1}$ are the Cross-Entropy loss functions of the target language of English and the learner's native language of Japanese, respectively. $I_{LID}(i)$ is the language ID indicator function that has the value 1 for speech sample i belonging to L2 and the value 0 for all training samples of L1.

Shared hidden layers can be seen as a feature extraction module which learns the commonality across English and Japanese based on their shared aspects, such as similar phones or the shared articulatory attributes. Acoustic model adaptation is consequently done during this training process. Non-native acoustic features extracted from the shared hidden layers are expected to provide better coverage of acoustic characteristics of the language learners. This architecture allows for learning non-native acoustic features without using a non-native dataset. Same as in Section III, one acoustic-phonetic model and four acoustic-articulatory models are trained. The acoustic-phonetic model is used in non-native speech recognition and pronunciation error detection components while the articulatory models are for diagnosing the pronunciation error at articulation level. We should note that the output layer for L1 will be removed during testing.

B. *Multi-task Knowledge Transfer*

We also explore the knowledge transfer effect of another related task. To fairly compare the effectiveness of different knowledge, a similar architecture to that used in cross-lingual transfer is adopted as shown in Fig. 5. There are two tasks of phonetic and articulatory modeling in this work. As introduced in Section III-A, there is a very close relationship between the phones and the articulatory attributes. For example, different vowels (/IH/ and /AX/) are mapped to a same attribute (unrounded). When we train the acoustic-phonetic model used for speech recognition, the knowledge of one kind of articulatory attribute is used. The manner of articulation is found to be the most useful knowledge, which we adopt in this study. When we train four articulatory models for pronunciation error diagnosis, we use the phonetic knowledge. Compared to the cross-lingual transfer, two differences should be noted. One is we only employ the L2 native data to train the acoustic model while two native language datasets are used in the cross-lingual transfer. The other difference is in the training process. In the cross-lingual transfer, only hidden layers are trained using samples of two languages. In other words, the output layer is separately trained with the samples of each language. As for those model parameters in Fig. 5, both hidden layers and two output layers are trained with all speech samples. The loss function is defined as:

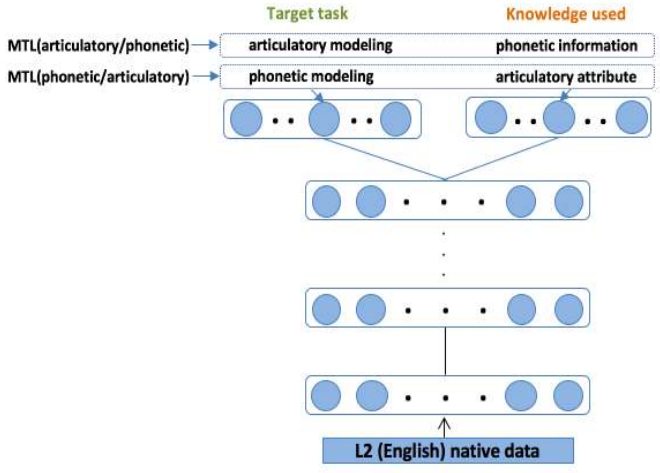


Fig. 5. Diagram of transfer learning based acoustic modeling complemented with an auxiliary task. By using equal weights, articulatory models and phonetic model are jointly trained using a multi-task learning (MTL) framework. The articulatory models are used for pronunciation error diagnosis while the phonetic model is used to recognize the non-native speech and detect the pronunciation errors.

$$loss_{TL(related\ task)} = loss_{T} + \omega * loss_{K} \quad (2)$$

where $loss_T$ and $loss_K$ are the Cross-Entropy loss functions of the target primary task and the related secondary task, respectively. In theory and conventions, these two loss terms can be weighted. In our previous study [11], we tuned the weight and found there is no significant performance difference among different weight values (placed on ω from 0.1 to 1.0). Based on this finding, we regard the two terms with a same weight in the above equation so that the model architecture essentially becomes a multi-task DNN.

V. Corpus and Experimental Setup

Three native speech corpora and a non-native speech corpus are used in this experiment. The native speech corpora are used to train the acoustic models and evaluate the performance of different acoustic modeling methods for native attribute recognition. The non-native testing corpus is to evaluate the performance on all three modules of the proposed CAPT system.

A. *Native Database*

The native corpus for L2 are Wall Street Journal (WSJ) database [64] and LibriSpeech database [65], which are commonly used for English large-vocabulary continuous speech recognition research. Sixty-four hour speech data from the SI-284 training data (WSJ0 and WSJ1) is selected after removing noisy utterances. There are 282 different speakers in total. It is used in both conventional DNN based acoustic modeling and two TL based approaches. Another sixty-four hour speech data from the LibriSpeech "train-clean-100" subset were used to investigate the effect of training data size. We conduct the native attribute recognition on two standard testing datasets of WSJ (Nov'92 and Nov'93). The other native corpus for L1 is JNAS [66], which is recorded by Japanese native speakers. It is also a commonly used database for Japanese large vocabulary continuous speech recognition research. We randomly select sixty-four hour speech utterances uttered by 324 speakers. This L1 native dataset is incorporated into the

cross-lingual based TL modeling which aims to characterize the phonological processes in Japanese speaking students learning English.

B. *Non-native Database*

The foreign language learners' speech database is a set of English words spoken by Japanese college students [67]. There are 7 speakers (2 male, 5 female) and each speaker uttered a same set of 850 basic English words. Each word contains phones which are difficult for Japanese students. These phones either do not exist in Japanese language or are pronounced in a very different manner. This non-native dataset is used to evaluate the performance of different modeling methods on the three tasks: non-native speech recognition, pronunciation error detection, and pronunciation error diagnosis.

C. *Experiment setup*

The acoustic feature consists of 40-dimensional log Mel-scale filter-bank outputs plus first and second temporal derivatives. The input to the network is 11 contiguous frames, 5 frames on each side of the current frame. The neural network has 7 hidden layers with 2048 nodes per layer. DNN training consists of unsupervised pre-training and supervised fine-tuning. All modeling methods in this paper adopt the same configuration above, and hyper-parameters are optimized on the development data set (Dev'93) of WSJ.

VI. Experiments and Analysis

In order to assess the performance of the acoustic modeling methods, we conduct experiments on the four tasks and present their results in this section. We first assess our acoustic models on attribute recognition of English native speakers for reference, before conducting evaluations on non-native English speech.

A. *Native Attribute Recognition Results*

The CAPT system should not only be effective for non-native foreign language learners but also work for native speakers whose pronunciation is regarded as the gold standard. We first assess our acoustic models on attribute recognition of English native speakers. Since the native training data are incorporated in all different model training processes and the testing datasets come from the same corpus, there is no mismatch problem in this experiment. We conduct free articulatory attribute recognition for the native speakers. Similar to the phone error rate, the attribute recognition error rate is used as an evaluation measure, which is computed over all four articulatory attributes.

We show the recognition error rates of all attributes in Fig. 6. Compared to DNN-64h, a further error reduction can be obtained when we increase the training data to 96 hours but no further improvement by increasing it to 128 hours. TL-128h (cross-lingual knowledge), which uses the same amount of data in total (128 hours), brings limited effect with mixed language datasets. On the other hand, the multi-task learning using phonetic knowledge achieves much more improvement. It substantially reduced the error rate on both "Nov'92 and

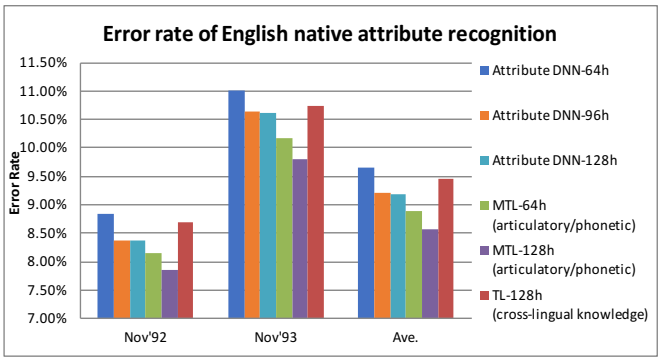


Fig. 6. Articulatory attribute recognition error rate of English native speakers on two standard testing datasets of WSJ corpus. The models are named after the total number of training data. The models marked with 64h are trained on the WSJ corpus only. Ave. means the average performance over two datasets.

Nov'93" testing datasets. When we average the performance over the two datasets, the recognition error rate is reduced from 9.64% to 8.89% with 64 hours of training data. The relative error reduction over the DNN-64h baseline is 7.82%, which is even better than the model trained with doubled data (DNN-128h). The effect is maintained for the model trained with 128-hour data (DNN-128h and MTL-128h). These results confirmed that phones and articulatory attributes are closely related, and phonetic knowledge help improve the performance of articulatory attribute recognition.

The articulatory attribute recognition method, in principle, can also be directly applied to non-native speakers for pronunciation error identification. However, the strategy of free decoding is expected to bring poor performance for non-native speech because of the significant difference between native and non-native speech. To support the CAPT system with attribute recognition of non-native speech, we restrict the recognition conducted on each detected phone error. Its details will be described in Section VI-D.

B. *Non-native Speech Recognition Results*

In the previous experiment, we have shown the proposed methods are effective for native speakers. From this subsection onwards, we focus on non-native speakers, which is our main target. The non-native speech corpus is used to evaluate the performance of different methods.

Accurate recognition of the learner's speech is important for the text-independent CAPT system. We evaluate the proposed methods on non-native speech recognition in this section. The knowledge from the articulatory attributes is used to improve acoustic-phonetic modeling. We conduct word recognition experiments with different settings. One is continuous speech recognition (CSR) while the other is isolated word recognition (IWR) which is more constrained. No language model is used in addition to the lexicon. Although all testing samples are single words, we do not assume the number of words in CSR so that both insertion and deletion errors could happen in the recognition result. As for the configuration in IWR experiment, word insertions and deletions are not allowed to occur because only one single word is assumed when decoding each utterance. To deal with pronunciation variation of non-native speakers,

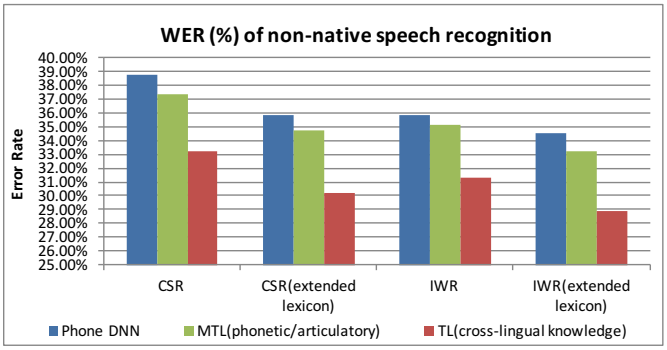


Fig. 7. English non-native word recognition on different settings. "CSR (extended lexicon)" means the continuous speech recognition is conducted using a lexicon in which each word has multiple pronunciation entries.

we also construct an extended lexicon, in which each word is represented by both canonical pronunciation and other possible pronunciation variations. These variations are derived by comparing the phone inventories of English and Japanese. In this section, word error rate (WER) is adopted as an evaluation metric.

Fig. 7 demonstrates the recognition performance of three different modeling methods at various conditions. We observe that TL based modeling methods consistently perform better than the DNN baseline under all four different recognition configurations. The main reason of high WERs is there are many minimal pairs (e.g. "frame" and "flame") included in the dataset. Confusing phones in these minimal pairs are useful for pronunciation learning while they are a challenge for speech recognition. The phone sequences of two words differ only in one phone so that misrecognizing a single phone will result in the whole word recognition error. Compared with the method that transfers articulatory knowledge to acoustic-phonetic modeling, the cross-lingual based TL method outperforms the DNN baseline by a large margin. The relative reduction of WER over the DNN baseline is 14.3% in CSR, 15.9% in CSR with the extended lexicon, 12.5% in IWR, and 16.2% in IWR with the extended lexicon. Compared with the result of native speech (Fig. 6), transferring knowledge of the learner's native language is more useful for non-native speech. The shared hidden layers are able to implicitly learn a better feature representation of the non-native acoustic space. In addition, when we employ the extended lexicon, even larger WER reduction is observed on both CSR and IWR tasks. The extended lexicon compensates for the pronunciation variants of non-native speakers, and this compensation works on the constructed decoding network, which is different from the acoustic model compensation by TL method. As a result, a synergetic effect is obtained when we combine these two methods together. The combination further decreases the WER to 30.16% which yields 22% relative improvement over the DNN baseline. In the IWR task, similarly, 19% relative WER reduction is achieved, which leads to the best WER of 28.91%.

C. *Pronunciation Error Detection Results*

As shown in Fig. 1, the two-pass pronunciation error detection and diagnosis framework is adopted in this study. The pronunciation error detection is conducted in the first pass which tells whether any phone is correct or not. We adopt the DNN-based GOP score proposed in [33] to detect the pronunciation errors. The GOP score of a target phone p given the observations $\mathbf{o}$ is computed as:

$$\mathrm{GOP}(p) \approx \log \frac{\mathrm{P}(p|\mathbf{o};\, t_s, t_e)}{max_{\{q\in Q\}}\mathrm{P}(q|\mathbf{o};\, t_s, t_e)} \tag{3}$$

where t_s and t_e are the start and end frame indexes of observations $\mathbf{o}$, which is obtained by Viterbi alignment of the ASR output in the text-independent scenario. In the evaluation of this section, however, we use the ground-truth text (that the learners should utter) for this GOP computation to make a meaningful comparison of the different methods. Q represents the whole phone sets. It is approximately calculated by the log posterior ratio between the target canonical phone p and its most competing phone q which has the highest posterior probability.

The log posterior of phone p given the observations $\mathbf{o}$ is computed as:

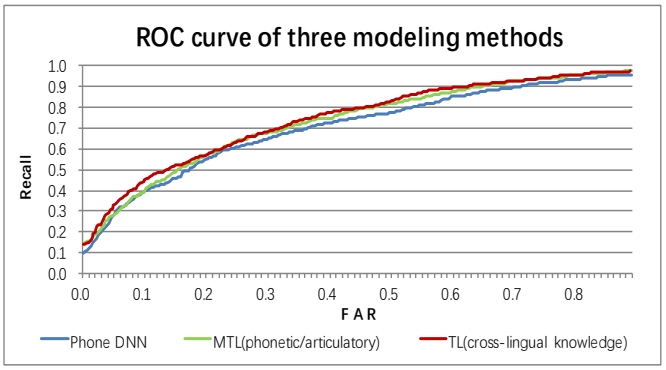
$$\log \mathrm{P}(p|\mathbf{o};\, t_s, t_e) \approx \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} \log \sum_{s\in p} \mathrm{P}(s|o_t) \tag{4}$$

where o_t is the input acoustic feature of frame t, s is the senone label belonging to the phone p. $P(s|o_t)$ is given by the softmax output of the senone DNN.

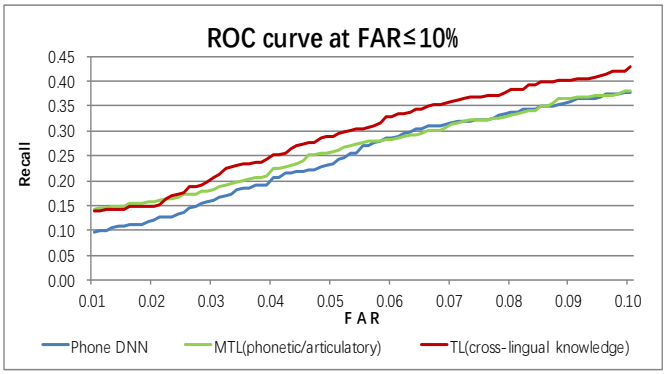
We calculate the GOP score for each phone segment using the equations defined above. A unique phone-independent threshold is then used to determine whether it is a pronunciation error or not.

To illustrate the detection performance of this binary pronunciation error classifier, receiver operating characteristic curve (ROC curve) is used. It is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold values. The TPR is also known as Recall and the FPR is also known as false alarm rate (FAR). In the task of pronunciation error detection, a false positive occurs when a learner's pronunciation is judged as incorrect, but actually is not a pronunciation error. FAR, as a result, shows how many incorrect results occur among learners' all correct pronunciations during the detection. The metric of Recall defines the fraction of true errors that are detected over the total number of pronunciation errors. We draw the ROC curve for three different methods in Fig. 8(a).

As shown in this figure, two TL based models outperform the baseline DNN system consistently when we vary the threshold for the GOP score. When we compare the effects of two different knowledge based TL methods, similar to that result observed in non-native speech recognition (Fig. 7), the knowledge derived from the learner's native language is more effective. This better performance mainly benefits from the more accurate forced alignment senone sequence of the non-native speech, which is generated from the adapted acoustic model. As there is an inherent trade-off between FAR and Recall, increasing the GOP threshold will result in a higher



(a)



(b)

Fig. 8. Performance of three modeling methods on pronunciation error detection. The best possible detection method would yield a point in the upper left corner or coordinate (0.0, 1.0) of the ROC space, representing a perfect classification, 100% accuracy (no false detection). A low FAR (less than 10%) is more interesting in this study and its zoomed-in view is shown in (b).

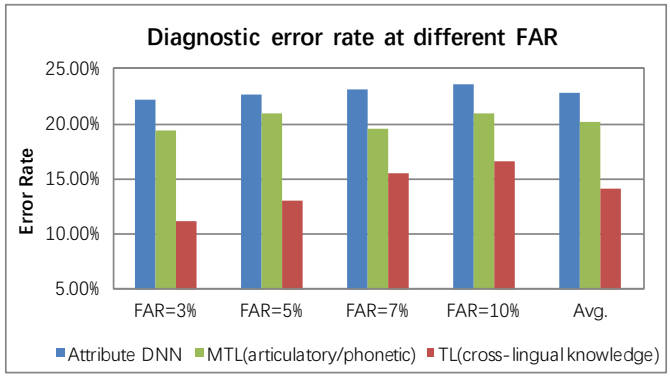


Fig. 9. Performance of three modeling methods on pronunciation error diagnosis. Avg. means the average DER over four different FAR settings.

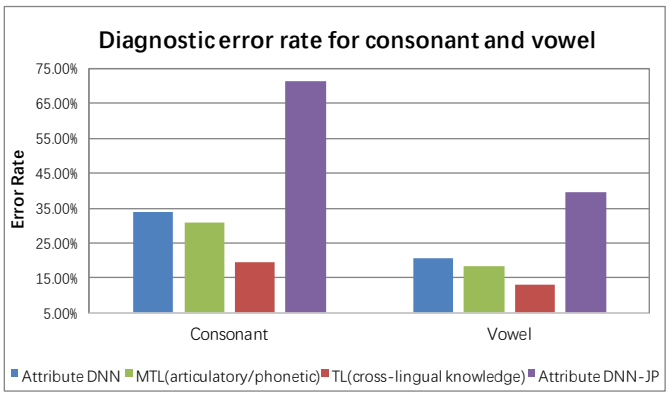


Fig. 10. Pronunciation error diagnostic performance of three modeling methods on consonant and vowel categories. It is averaged over the four FAR settings shown in Fig. 9. The model marked with JP is trained on the JNAS corpus only.

risk of FAR despite of a better coverage of the true pronunciation errors. Considering the purpose of CAPT, a high FAR is not desirable because it would discourage learners by rejecting their correct pronunciation. In this study, we control the risk of FAR below 10%. Fig. 8(b) shows a zoomed-in view of ROC curve where FAR is less than or equal to 10%. At the operating point where FAR equals 10%, the recall of true errors is increased from 37.8% to 42.8% by the cross-lingual based TL.

D. Pronunciation Error Diagnosis Results

Pronunciation error diagnosis serves as the last but the most critical part in the CAPT system. It identifies the detected errors in a higher resolution and is conducted based on the results of the first pass of pronunciation error detection. During the error detection pass, the system points out which part of learner's speech is probably a pronunciation error and should be considered in the diagnosis pass. The non-native attribute recognition as a consequence becomes easier now because it only needs to focus on a very short segment. This scenario is different from the native attribute recognition on the whole utterance without any constraint. Articulatory attributes are focused to characterize the segmental pronunciation errors. The attributes are related to human speech production so that articulator related feedbacks could be naturally generated by comparing recognized attributes with the canonical attributes. We conduct the articulatory attribute recognition in the detected segments. Diagnostic error rate (DER) [68] is adopted as an evaluation measure.

The diagnostic performance at different thresholds on the GOP score for controlling the FAR is plotted in Fig. 9. The threshold is adjusted so that a reasonable low risk of FAR which equals 3%, 5%, 7%, or 10% is adopted. From the figure, we see that in all different settings, TL based methods perform better than the DNN baseline. Cross-lingual knowledge based TL achieves a lower DER than the phonetic knowledge transfer, in which phonetic information is used as the source knowledge to improve the articulatory modeling. This result shows effectiveness of the cross-lingual transfer for the articulatory modeling.

For all three models, as the FAR gradually increased from 3% to 10%, the DER begins to rise as well. This increase is caused by the thresholds adopted in the different settings of FAR. If a looser threshold is adopted, more pronunciation samples being less confident of an error will be included. These less confident target samples will pose a bigger challenge to the system. We present the average performance over those four FAR settings at the rightmost in Fig. 9. In contrast to the 2% improvement achieved in the phonetic knowledge based TL, a

much larger DER decrease from 22.8% to 14.1% is observed with cross-lingual TL.

Fig. 10 gives the diagnostic performance breakdown for consonants and vowels. From the result, we first observe vowels are better identified than consonants. This is mainly because there are fewer articulatory attribute classes in the vowel category. Considering the linguistic interference, we also conduct the attribute recognition using Japanese acoustic-articulatory models, which are trained with the JNAS corpus only. From the diagnostic results shown in Fig. 10, we see the error rate is almost twice as much as the result with the English acoustic model. This is not surprising because of the significant phonological difference between Japanese and English.

When we review all experiments, we note the different tendencies of two kinds of knowledge presented on the three non-native speech related tasks and the native attribute recognition task. The related task knowledge is more helpful in native attribute recognition. However, the cross-lingual knowledge derived from the L1-L2 pair is more effective for non-native speech related tasks while no significant positive effect on native speech. This indicates that "what to transfer?" is important and should be carefully selected when applying transfer learning. In this study, transferring knowledge of the learner's native language is most relevant to cover characteristics of non-native speech.

VII. Conclusions

In this paper, we have proposed an effective acoustic model training for the CAPT application. Considering there exists very limited amount of non-native speech data, we address acoustic modeling without using target non-native speech in this paper. To mitigate the mismatch between native and non-native speakers, we characterize the non-native speech by exploiting two large native speech corpora of the learner's native language and the target foreign language based on the cross-lingual language transfer learning. The proposed DNN consists of shared hidden layers and language-dependent output layers to learn a better feature representation of non-native speech through the shared layers.

In the non-native speech recognition and pronunciation error detection experiments, we confirmed the effectiveness of the proposed cross-lingual based transfer learning on acoustic-phonetic modeling. For corrective feedback, we conduct the articulatory level diagnosis for each detected phone error. Experimental results demonstrate that the proposed method of acoustic-articulatory modeling based on cross-lingual transfer learning is effective.

Though allowing the language learners to speak freely is desired, it brings challenges as well. First, the system needs to recognize every intended word even if the learner mispronounced it to another word. It is also challenging to get an accurate alignment boundary when the learner's pronunciation is incorrect. Incorporating appropriate linguistic models and contextual models will be beneficial for speech recognition while adding possible pronunciation errors to the alignment graph might be helpful for forced alignment. These directions should be explored in the future.

In this paper, we have demonstrated the effect of transferring the knowledge from learner's native language with a fully connected feedforward network architecture. In our future work, more sophisticated deep learning models such as neural attention models will be investigated. The proposed method, in theory, can be applied to any language pairs as long as there is a native corpus. We have applied our proposed method to English language learning. In the future, we will apply it to more language learning datasets.

References

[1] Qian X, Meng H, Soong F, "A two-pass framework of mispronunciation detection and diagnosis for computer-aided pronunciation training," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 24, no. 6, pp. 1020-1028, Jun. 2016.

[2] C.-H. Jo, T. Kawahara, S. Doshita, and M. Dantsuji, "Automatic Pronunciation Error Detection and Guidance for Foreign Language Learning," in Proc. Int. Conf. Spoken Lang. (ICSLP), pp.2639–2642, 1998.

[3] J. C. Koreman, P. Wik, O. Husby, and E. Albertsen, "Universal contrastive analysis as a learning principle in CAPT," in Workshop on Speech and Language Technology in Education (SLaTE), pp. 172–177, 2013.

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

[5] V. Arora, A. Lahiri, H. Reetz, "Phonological feature-based speech recognition system for pronunciation training in non-native language learning," in The Journal of the Acoustical Society of America, vol. 143, no. 1, pp. 98-108, Jun.2018

[6] A. Rathinavelu, H. Thiagarajan, and A. Rajkumar, "Three dimensional articulator model for speech acquisition by children with hearing loss," in Proc. the 4th Int. Conf. Universal Access in Human Computer Interaction, pp. 786-794, 2007.

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

[8] 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 Commun, vol. 52, pp. 493-503, 2010.

[9] N. F. Chen, D. Wee, R. Tong, B. Ma, and H. Li,"Large-Scale Characterization of Non-Native Mandarin Chinese Spoken by Speakers of European Origin: An Analysis on iCALL," in Speech Commun, vol. 84, pp. 46-56, 2016.

[10] R. Duan, T. Kawahara, M. Dantsuji, and J. Zhang, "Effective articulatory modeling for pronunciation error detection of L2 learner without non-native training data," in Proc. IEEE Int. Conf. Acoust. Speech. Signal Process (ICASSP), pp. 5815-5819, 2017.

[11] R. Duan, T. Kawahara, et al, "Articulatory modeling for pronunciation error detection without non-native training data based on DNN transfer learning", IEICE pp.2174-2182, 2017.

[12] J. Huang, J. Li, D. Yu, L. Deng, and Y. Gong. "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers." In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 7304-7308, 2013.

[13] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M.A. Ranzato, M. Devin, and J. Dean. "Multilingual acoustic models using distributed deep neural networks." In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 8619-8623, 2013.

[14] S.Stefano, P. Laface, L. Fissore, R. Gemello, and F. Mana. "On the use of a multilingual neural network front-end." In Ninth Annual Conference of the International Speech Communication Association, 2008.

[15] K. Veselý, M. Karafiát, F. Grézl, M. Janda, and E. Egorova. "The language-independent bottleneck features." In 2012 IEEE Spoken Language Technology Workshop (SLT), pp. 336-341, 2012.

[16] N. Moustroufas and V. Digalakis, "Automatic pronunciation evaluation of foreign speakers using unknown text," In Computer Speech and Language, vol. 21, no. 6, pp. 219–230, 2007.

[17] L. Chen, K. Zechner, and X. Xi, “Improved pronunciation features for construct-driven assessment of nonnative spontaneous speech,” In Proc. NAACL, 2009.
[18] H. Meng, Y.Y. Lo, L. Wang, and W.Y. Lau, “Deriving salient learners’ mispronunciations from cross-language phonological comparisons,” in Proc. ASRU, pp.437–442, 2007.
[19] H. Strik, K. Truong, F. De Wet, and C. Cucchiarini, “Comparing different approaches for automatic pronunciation error detection,” in Proc. Speech Communication, vol.51, pp.845–852, 2009.
[20] 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. IEEE Int. Conf. Acoust. Speech. Signal Process (ICASSP), pp.8232– 8236, 2013.
[21] A. Lee and J. Glass, “Mispronunciation Detection without Nonnative Training Data,” in Proc. Interspeech, pp.643–647, 2015.
[22] S. Joshi, N. Deo, and P. Rao, “Vowel mispronunciation detection using DNN acoustic models with cross-lingual training,” in Proc. Interspeech, pp.697–701, 2015
[23] Z. Wang, T. Schultz, and A. Waibel, “Comparison of acoustic model adaptation techniques on non-native speech,” In 2003 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 540–543, 2003.
[24] G. Bouselmi, D. Fohr, I. Illina, and J.-P. Haton, “Multilingual non-native speech recognition using phonetic confusion based acoustic model modification and graphemic constraints,” in Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP ’06), vol. 1, pp. 109–112, 2006.
[25] U. Uebler, M. Boros, “Recognition of Non-native German Speech with Multilingual Recognizers,” in Proc. Eurospeech, pp. 911-914, 1999.
[26] M. Matassoni, R. Gretter, D. Falavigna, and D. Giuliani, “Non-native children speech recognition through transfer learning,” in Proc. IEEE Int. Conf. Acoust. Speech. Signal Process. (ICASSP), pp. 6229-6233, 2018.
[27] S. Schaden, "Generating Non-native Pronunciation Lexicons by Phonological Rule," In Proc. Int. Conf. Spoken Lang (ICSLP), pp. 2545-2548, 2004.
[28] Y. R. Oh, J. S. Yoon, and H. K. Kim, “Adaptation based on pronunciation variability analysis for non-native speech recognition,” in Proc. IEEE Int. Conf. Acoust. Speech. Signal Process. (ICASSP), 2006, pp. 137–140.
[29] Y. Kim, H. Franco, and L. Neumeyer, “Automatic pronunciation scoring of specific phone segments for language instruction”, in Proc. Eurospeech, 1997.
[30] F. Zhang, C. Huang, F. Soong, M Chu, R. Wang, “Automatic mispronunciation detection for Mandarin,” in Proc. IEEE Int. Conf. Acoust. Speech. Signal Process. (ICASSP), 2008, pp. 5077–5080
[31] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, “Automatic detection of phone-level mispronunciation for language learning,”, in Proc. Eurospeech, pp. 851-854, 1999
[32] S. M. Witt and S. J. Young, “Phone-level pronunciation scoring and assessment for interactive language learning,” Speech Communication., vol. 30, no. 2, pp. 95–108, 2000.
[33] W. Hu, Y. Qian, F. 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.
[34] W. Hu and F.K. Soong, “KL-divergence based mispronunciation detection via DNN and decision tree in the phonetic space," in Proc. Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016.
[35] I. Amdal, M. H. Johnsen, and E. Versvik. Automatic evaluation of quantity contrast in nonnative Norwegian speech. In Proc. SLaTE, 2009.
[36] H. Strik, K. Truong, F. De Wet, and C. Cucchiarini, “Comparing different approaches for automatic pronunciation error detection,” in Proc. Speech Communication, vol.51, pp.845–852, 2009
[37] K. Truong, A. Neri, C. Cucchiarini, and H. Strik, “Automatic pronunciation error detection: an acoustic-phonetic approach,” in InSTIL/ICALL Symposium 2004, 2004.
[38] A. M. Harrison, W. K. Lo, X. Qian, and H. Meng, “Implementation of an Extended Recognition Network for Mispronunciation Detection and Diagnosis in Computer-Assisted Pronunciation Training,” In Proc. Interspeech, 2009.
[39] Y. Tsubota, T. Kawahara, M. Dantsuji, “Recognition and verification of English by Japanese students for computer-assisted language learning system,” In Proc. ICSLP, pp. 1205–1280, 2002.
[40] J. Kim, C. Wang, M. Peabody, and S. Seneff, “An interactive English pronunciation dictionary for Korean learners,” In Proc. Interspeech, 2004.
[41] A. M. Harrison, W. Y. Lau, H. Meng, and L. Wang, “Improving mispronunciation detection and diagnosis of learners’ speech with context-sensitive phonological rules based on language transfer,” In Proc. Interspeech, 2008.
[42] W. K. Lo, S. Zhang, and H. Meng, “Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system,” In Proc. Interspeech, 2010.
[43] A. Lee and J. Glass, “Mispronunciation detection without nonnative training data,” in Proc. Interspeech, 2015.
[44] J.R. Westbury, “X-ray Microbeam Speech Production Database User’s Handbook”, Waisman Center on Mental Retardation and Human Development. University of Wisconsin, Madison, WI, USA, version 1.0 edition. 1994
[45] AA. Wrench, “Multi-channel/multi-speaker articulatory database for continuous speech recognition research”, Phonus 5, 1–13, 2000.
[46] S. Narayanan, K. Nayak, S. Lee, A. Sethy, D. Byrd, “An approach to real-time magnetic resonance imaging for speech production”, in J. Acoust. Soc. Am. Vol. 115, pp. 1771–1776, 2004.
[47] M. Grimaldi, B.F. Gili, F. Sigona, M. Tavella, P. Fitzpatrick, L. Craighero, L. Fadiga, G. Sandini, and G. Metta, “New technologies for simultaneous acquisition of speech articulatory data: 3D articulograph, ultrasound and electroglottograph”, In Proc. LangTech, Italy, 2008.
[48] H. Li, J. Tao, M. Yang, B. Liu, “Estimate articulatory MRI series from acoustic signal using deep architecture”, in Proc. IEEE Int. Conf. Acoust. Speech. Signal Process. (ICASSP), 2015.
[49] B. Abraham, S. Umesh, “An automated technique to generate phone-to-articulatory label mapping”, in Speech Communication, vol 86, pp. 107–120, 2017.
[50] H. Zheng, Z. Yang, L Qiao, J. Li, W. Liu, “Attribute Knowledge Integration for Speech Recognition Based on Multi-task Learning Neural Networks”, in Proc. INTERSPEECH, 2015
[51] W. Li, S. Marco, Nancy F. Chen, and C-H Lee, “Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling”, in Proc. IEEE Int. Conf. Acoust. Speech. Signal Process. (ICASSP), 2016.
[52] Siniscalchi, Sabato Marco, Dau-Cheng Lyu, Torbjørn Svendsen, and Chin-Hui Lee. "Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data." IEEE transactions on audio, speech, and language processing 20, no. 3 (2012): 875-887.
[53] Yu, Dong, Sabato Marco Siniscalchi, Li Deng, and Chin-Hui Lee. "Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition." In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp. 4169-4172. IEEE, 2012.
[54] G. Hinton, L. Deng, D. Yu, GE. Dahl, AR. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, TN. Sainath, B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” in IEEE Signal Processing Magazine, vol. 29 no. 6, pp.82-97, Nov, 2012.
[55] Matthew E Taylor and Peter Stone, “Transfer learning for reinforcement learning domains: A survey,” in Journal of Machine Learning Research, vol. 10, pp. 1633–1685, 2009.
[56] Sinno Jialin Pan and Qiang Yang, “A survey on transfer learning,” in IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, 2010.
[57] Yoshua Bengio, “Deep learning of representations for unsupervised and transfer learning,” in Proc. ICML Unsupervised and Transfer Learning, pp.17-36, 2012.
[58] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, “Transfer learning using computational intelligence: A survey,” in Proc. Knowledge-Based Systems, vol. 80, pp. 14–23, 2015.
[59] J. Huang, J. Li, D. Yu, L. Deng, and Y. Gong. "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers." In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 7304-7308. IEEE, 2013.
[60] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M.A. Ranzato, M. Devin, and J. Dean. "Multilingual acoustic models using distributed deep

neural networks." In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 8619-8623. IEEE, 2013.
[61] L. Postman and K. Stark, “Role of response availability in transfer and interference,” in Journal of Experimental Psychology, vol. 79, no. 1, 1969.
[62] J.D. Bransford, A.L. Brown, and R.R. Cocking, “How people learn: Brain, mind, experience, and school,” Wasington, DC, National Academy Press, 1999.
[63] C.B. Chang and A. Mishler, “Evidence for language transfer leading to a perceptual advantage for non-native listeners,” in Journal of the Acoustical Society of America, vol.132, no.4, pp.2700–2710, 2012.
[64] D.B. Paul and J.M. Baker, “The design for the Wall Street Journal-based CSR corpus,” in Proc. of the workshop on Speech and Natural Language. Association for Computational Linguistics, pp. 357–362, 1992.
[65] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an ASR corpus based on public domain audio books,” In Acoustics, Speech and Signal Processing (ICASSP), pp. 5206-5210, 2015.
[66] 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, no.3, pp.199–206.
[67] K. Tanaka, H. Kojima, Y. Tomiyama, and M. Dantsuji, “Acoustic models of languageindependent phonetic code systems for speech processing,” in Proc. Spring Meeting of the Acoustical Society of Japan, pp. 191–192, 2001.
[68] K. Li, X. Qian, H. Meng, “Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks,” in IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 25, no. 1, pp. 193-207, 2017.

Richeng Duan received the B.E. degree in Computer Science from Tianjin University of Technology, China, in 2011, M.S. degree in Information Science from Beijing Language and Culture University in 2015, and Ph.D. in Informatics from Kyoto University, Japan, in 2018. His research interests include speech recognition, model adaptation, and computer-assisted language learning. He is currently a researcher at I2R, A*STAR, Singapore.

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 400 academic papers on speech recognition, spoken language processing, and spoken dialogue systems. He has been conducting several projects including speech recognition software Julius, the automatic transcription system for the Japanese Parliament (Diet), and autonomous android ERICA. 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 was an editorial board member of Elsevier Journal of Computer Speech and Language and IEEE/ACM Transactions on Audio, Speech, and Language Processing. He is the Editor-in-Chief of APSIPA Transactions on Signal and Information Processing. Dr. Kawahara is a board member of APSIPA and ISCA, 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.

Hiroaki NANJO received the B.E. degree in 1999, the M.E. degree in 2001, and the Ph.D. degree in 2004 from Kyoto University, Kyoto, Japan. During 2004 to 2007, he was a Research Associate, and during 2007 to 2015, he was an Assistant Professor at Department of Media Informatics, Faculty of Science and Technology, Ryukoku University. From 2015, he is an Associate Professor at Academic Center for Computing and Media Studies, Kyoto University. He has been working on speech recognition and understanding. He is a member of Acoustical Society of Japan (ASJ), Information Processing Society of Japan (IPSJ), Institute of Electronics, Information and Communication Engineers (IEICE), Institute of Electrical and Electronics Engineers (IEEE), The Virtual Reality Society of Japan (VRSJ), and The Japan Association for Language Education and Technology (LET).