Transfer Learning based Non-native Acoustic Modeling for Pronunciation Error Detection

Richeng Duan[1], *Tatsuya Kawahara*[1], *Masatake Dantsuji*[2], *Hiroaki Nanjo*[2]

[1]School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
[2]Academic Center for Computing and Media Studies, Kyoto University
duan@sap.ist.i.kyoto-u.ac.jp

Abstract

The scarcity of large-scale non-native corpora and human annotations are two fundamental challenges in the development of computer-assisted pronunciation training (CAPT) systems. We explored several transfer learning based methods to detect the pronunciation errors without using non-native training data. Effects were confirmed in the Mandarin Chinese pronunciation error detection of Japanese speakers. In this paper, we investigate the generality of the methods through application to an English speech data of Japanese speakers. We also evaluate on a non-native phone recognition experiment, which is necessary but challenging in advanced CAPT systems. Experimental results show that transfer learning based acoustic modeling methods can not only be ported to a new target language but also effective in a recognition task.

Index Terms: computer-assisted pronunciation training (CAPT), language independent pronunciation error detection, transfer learning

1. Introduction

In recent years, automatic speech recognition (ASR) has achieved great progress due to the emergence of Deep Neural network (DNN) and big data. However, pronunciation error detection in CAPT systems, which is often based on ASR technology, cannot benefit a lot from the ASR success. One of the most important reasons is lacking large-scale training resources with qualified annotations. As pronunciation of a target language is easily affected by learners' native language, it is better to train the acoustic model with learners' speech data. Unfortunately, it is much more challenging in collecting and labeling non-native speech than native speech because of the fewer populations and unnatural pronunciations [1]. To overcome this problem, we have explored several transfer learning based methods which aim at effective learning DNN acoustic models of non-native speakers without using such training data [2].

In this paper, we investigate the generality of transfer learning based methods in terms of different target languages and different evaluation settings. According to the Ethnologue [3], there are around 7000 living languages in the world. It is impossible to collect and annotate every non-native speech corpus considering different language pairs (target language and learners' mother tongue). Most existing approaches to non-native acoustic modeling in CAPT are language dependent while our proposed approaches do not make any assumption about language pairs. As a result, we are interested in investigating the generality when applying to other language learning corpora. Specifically, a non-native English speech corpus of Japanese students is used in this work (different from non-native Mandarin Chinese in [2]). Experimental results show that transfer learning based modeling methods can generalize well on this new target language learning corpus. The majority of pronunciation error detection works are based on read speech, which means students have to repeat the learning material. Providing more choices for learners is desirable for advanced language learning [4]-[6]. However, allowing more freedom means the system needs to recognize the learners' speech firstly, which is a challenging task in ASR because of a broad range of pronunciation variations in learners' speech. Here, we also investigate the effectiveness of transfer learning based methods when evaluated in the non-native speech recognition.

The rest of this paper is organized as follows: In Section 2, pronunciation error detection based on DNN articulation models is described. Section 3 presents transfer learning based methods to enhance the learning of the DNN models. The performance of these modeling methods is firstly evaluated on native articulatory attributes recognition task in Section 4. Section 5 and 6 confirm the effectiveness in pronunciation error detection experiment and speech recognition of non-native speakers. Conclusions are in the final section.

2. Pronunciation error detection based on DNN articulation model

2.1. Pronunciation error detection

Pronunciation error detection on segmental level has been a core component in CAPT system. Most of the prior works focused on detecting phone substitution errors. Some researchers targeted a few specific problematic phones and explore the distinctive features and classifiers [7]-[9]. Others conducted the detection based on ASR technology, either incorporating the possible errors into the pronunciation lexicon or directly adding them into the decoding grammar [10]-[16]. The ASR-based method is more general than the specially designed ones since it can detect all the phones in a unified framework. A typical feedback of phone error detection approach 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 [17]-[19]. Facing the same pronunciation error, 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 helpful in many areas, such as speech comprehension improvement [20], speech therapy [21] and pronunciation perceptual training [22].

2.2. 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 these articulatory attributes of foreign language learners.

Considering the co-articulation effect, context-dependent tri-attribute modeling is employed. 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. Since the mapping relation between articulatory attributes and phones is many-to-many, we prepare four kinds of transcriptions (manner, place-roundedness, place-backness and place-height) to represent all articulatory attributes. Articulatory attribute transcription is derived from the phone transcription. We exploit native data of target language (English in this work) to train the articulatory models (see Figure 1). These models can be directly used to detect pronunciation errors of language learners as a baseline.

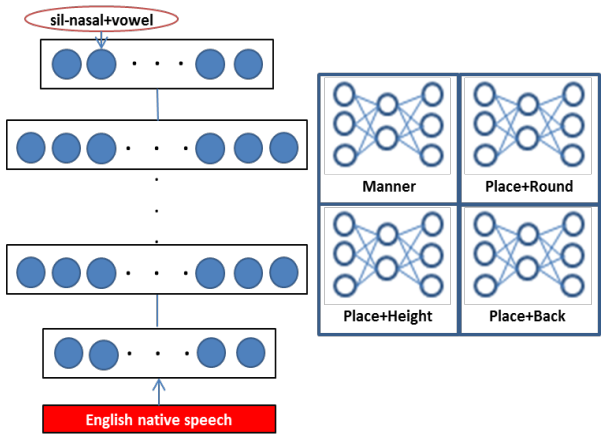

Figure 1: *Context-dependent modeling of articulatory attributes.*

3. Enhancing articulatory attribute modeling with transfer learning

The idea of transfer learning, which should trace back to 20 years ago, has been successfully employed in broad research fields [23]-[26]. We employ transfer learning on the articulation modeling of non-native speech. Inter-language transfer learning, related-task transfer learning, and combination of these two methods are explored.

3.1. Related-task transfer 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. In this study, phone classification task is served as the secondary task, which aims at helping the primary task learn better feature representation of attributes with the phonetic information.

3.2. Inter-language transfer learning on articulatory attribute modeling

In inter-language transfer learning method, two large native speech corpora of learners' native language (Japanese) and a target language (English) 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. Shared hidden layers in multi-lingual DNN (ML-DNN) allow for learning non-native articulatory features without using such data set.

3.3. Combining related-task and inter-language transfer learning for articulatory attribute modeling

We further investigate the combination of the related-task and inter-language transfer learning (see Figure 2). The related-task transfer learning learns the commonality through co-supervision of different tasks. The inter-language transfer learning aims at learning a better feature representation of non-native speech. As a result, their combination can have a synergetic effect.

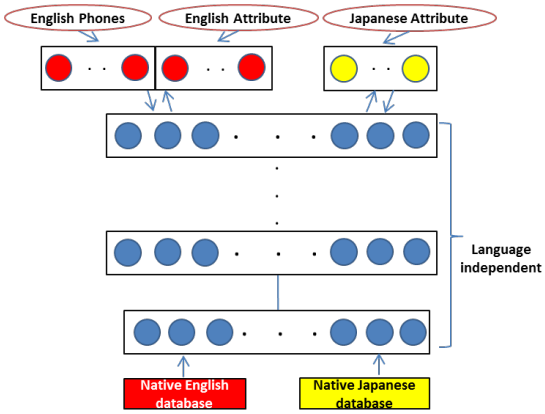

Figure 2: Enhancing the articulatory models through related-task and inter-language transfer learning.

4. Native Attribute Recognition Experiment

4.1. Database

Two native speech corpora are used in this experiment. The native English corpus is Wall Street Journal (WSJ) databases [27], which is used to train the target articulatory models and validate different modeling methods. The Japanese corpus named JNAS [28] is also a commonly used database for Japanese large-vocabulary continuous speech recognition research. Sixty-four hours speech data from each corpus were selected after filtering noisy utterances. We conduct the evaluation on both Nov'92 and Nov'93 data sets of WSJ.

4.2. System Configuration

All different methods use the following DNN configuration: the acoustic feature consists of 40-dimensional filter bank outputs plus their first and second temporal derivatives. The input to the network is made by splicing 11 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.

4.3. Experimental Results

The experimental results of different articulatory attributes are shown in Figure 3 to Figure 6. From these 4 figures, we observe the effects of all three transfer learning based methods. Compared with the conventional DNN, all the methods achieve lower recognition error rates. ML-DNN could benefit from more training data than conventional DNN though it comes from another language. MT-DNN is more effective than ML-DNN for native speech because it explicitly takes advantage of more direct information (phonetic labels). We highlight the effect of their combination (ML+MT DNN) as they are complementary to each other and can further reduce the error rate.

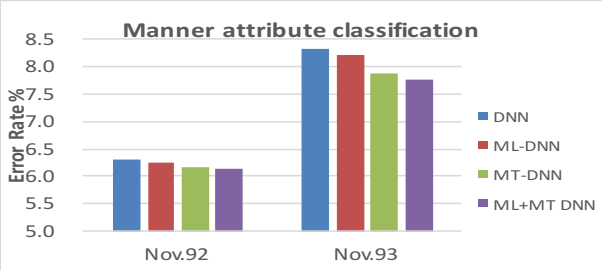

Figure 3: Manner attribute recognition.

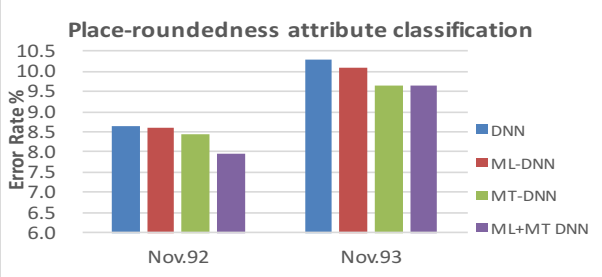

Figure 4: Place-roundedness attribute recognition.

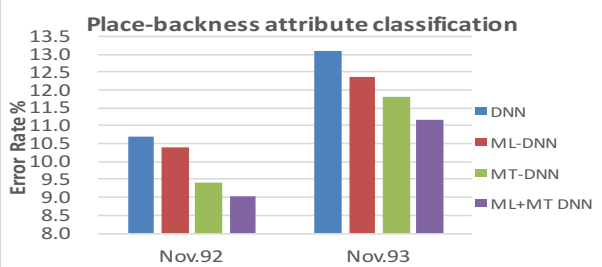

Figure 5: Place-backness attribute recognition.

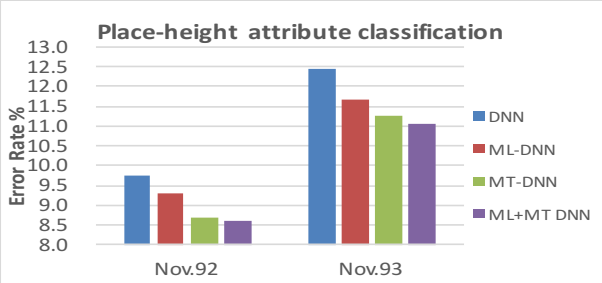

Figure 6: Place-height attribute recognition.

5. Non-native Pronunciation Error Detection

5.1. Experiment Setup

The evaluation data for pronunciation error detection is a corpus of English words spoken by Japanese students [29]. There are 7 speakers (2 male, 5 female) and each speaker uttered a same set of 850 English words. The database contains phonemic hand-labels, which were transcribed faithfully. We employ finite state decoding network for pronunciation error detection, which includes the canonical pronunciation and possible pronunciation errors.

5.2. Pronunciation Error Types

In this experiment, three pronunciation error types are focused which involve 5 specific vowels:

- Lip shape error: vowels with spread lips have problems of rounded sound.
- Tongue position error (horizontal): inappropriate tongue position with a little front or back.
- Tongue position error (vertical): inappropriate tongue position with a little high or low.

5.3. Evaluation Metrics

Two common used metrics of Detection Accuracy (DA) and F-score are used to evaluate the detection performance of different methods:.

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

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

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

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

N_{TE} is the number of pronunciation errors correctly detected 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.

5.4. Experimental Results

Figure 7 compares the overall detection performance of different methods: conventional DNN, MT-DNN, ML-DNN, the combined ML+MT DNN. We can see that both MT-DNN and ML-DNN perform better than the conventional DNN. 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 English data only while we add Japanese characteristics by using both English and Japanese data sets. Detailed detection results of individual error types are shown from Figure 8 to Figure 10. Among these three errors, the system detects the lip shape error best, while the tongue position error types are less accurate. This

tendency is similar to what we observed in the native speech attribute recognition experiment. However, the absolute performance of tongue backness error detection is rather low compared with the high accuracy of native place-backness attributes (Figure 5). This is partly due to the significant phonological differences between Japanese and English, especially the vowel system. In terms of vowel inventory, there are only five vowels in Japanese language while sixteen vowels (including the schwa sound) are in English. The considerably more vowels in English not only brings a big challenge for Japanese students learning English vowels but also for the annotators when labeling the non-native speech.

We also add the result of Mandarin Chinese pronunciation error detection [2] in Figure 11. We can see that transfer learning based methods can generalize well in different target languages. Compared with Mandarin Chinese pronunciation error detection, however, detection performance of English pronunciation error is much lower. This is because there are only vowel errors considered in this preliminary study, and the acoustic difference among English vowels is very subtler as discussed above.

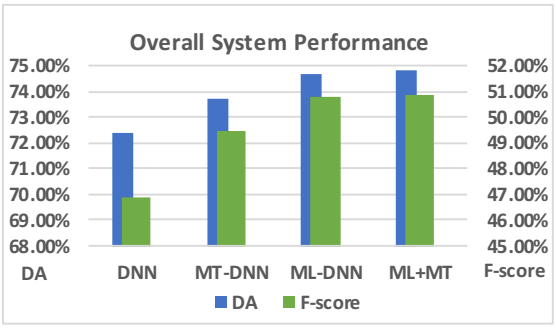

Figure 7: Overall system performance for non-native English pronunciation error detection.

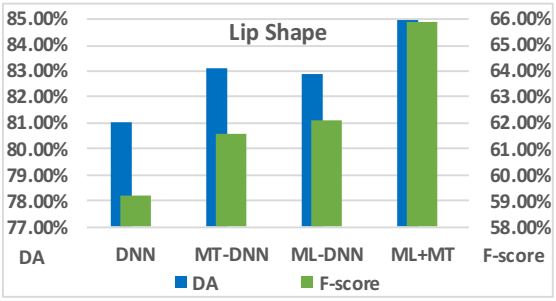

Figure 8: DA & F-score for lip shape error.

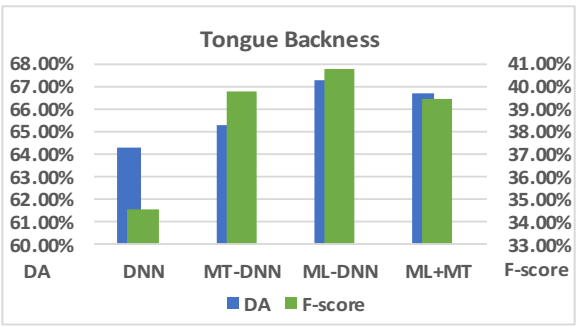

Figure 9: DA & F-score for tongue backness error.

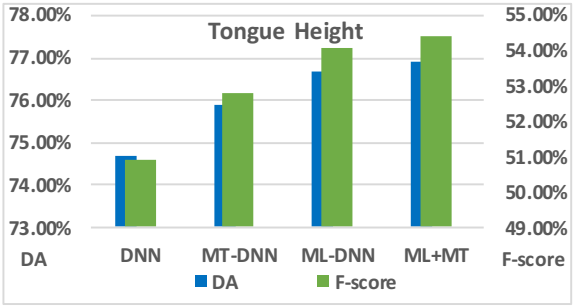

Figure 10: DA & F-score for tongue height error.

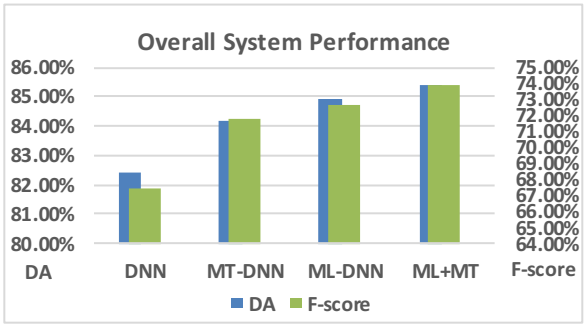

Figure 11: Overall system performance for non-native Mandarin Chinese pronunciation error detection.

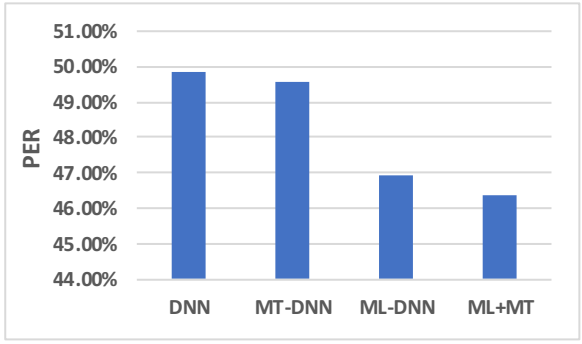

Figure 12: Phone error rate (PER) of non-native English speech recognition.

6. Non-native Speech Recognition

Above-mentioned pronunciation error detection is text-dependent, which means the system provides a set of pre-defined scripts for the students to read. In this case, the system knows what the target pronunciations should be, and the scripts can be used for constructing a finite state decoding network for pronunciation error detection. However, as the students improve, especially for those advanced learners, it would be better to let them speak freely and create their own sentences rather than reading a given text. The text-independent system should provide two functions. One is recognizing the learner's speech, the other is detecting the pronunciation errors based on recognized text.

We conduct a free phone recognition experiment, in which the system recognizes the non-native speech without using any constraint. This setting is the most general but hardest condition. Figure 12 demonstrates the phone error rate (PER) of different methods. We can see that transfer learning based

methods perform better than conventional DNN even when they are ported to speech recognition of non-native speech. We also see that there is a large room to improve the performance as the absolute recognition error rate is high. A promising solution is that adding a lexicon with reasonable size to control the search space.

7. Conclusions

In this paper, we investigate the generality of transfer learning based modeling methods by applying to different target languages and different tasks. We have performed pronunciation error detection experiments on two different language learning corpora (Mandarin Chinese and English). Experimental results show the effectiveness of transfer learning based modeling methods on both corpora. We then investigate the usability in speech recognition of non-native speech. The experimental result shows that transfer learning based methods are still effective even with the free phone recognition.

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 to more language learning corpora. Text-independent pronunciation error detection will be our next direction.

8. Acknowledgements

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

9. References

[1] 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," Speech Communication, 2016.

[2] R. Duan, T. Kawahara, M. Dantsuji, and J. Zhang, "Effective articulatory modeling for pronunciation error detection of L2 learner without non-native training data.," Proc. ICASSP , pp. 5815-5819, 2017

[3] Ethnologue: Languages of the World. Dallas, Texas: SIL International. Retrieved from http://www.ethnologue.com/16

[4] C. Molina, N. B. Yoma, J. Wuth, and H. Vivanco. ASR based pronunciation evaluation with automatically generated competing vocabulary and classifier fusion. Speech communication, 2009.

[5] Derrick Higgins, Xiaoming Xi, Klaus Zechner, and David Williamson, "A three-stage approach to the automated scoring of spontaneous spoken responses," Computer Speech & Language, vol. 25, no. 2, pp. 282–306, 2011.

[6] Automated Scoring of Speech, http://www.ets.org/research/topics/as nlp/speech, last accessed, September 27, 2016.

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

[8] H. Strik, K. Truong, F. De Wet, and C. Cucchiarini, "Comparing classifiers for pronunciation error detection," Proc. Interspeech, pp. 1837-1840, 2007.

[9] 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.

[10] 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.

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

[12] Y.-B. Wang, 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.

[13] 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, 2013.

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

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

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

[17] J. Tepperman and S. Narayanan, "Using articulatory representations to detect segmental errors in nonnative pronunciation," in IEEE Transactions on Audio, Speech, and Language Processing, 16(1):8-22, Jan. 2008.

[18] O. Engwall, "Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher," Computer Assisted Language Learning, 25:37–64, 2012.

[19] 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

[20] 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.

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

[22] 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.

[23] Matthew E Taylor and Peter Stone, "Transfer learning for reinforcement learning domains: A survey," Proc. Journal of Machine Learning Research, vol. 10, pp. 1633–1685, 2009.

[24] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," Proc. IEEE Transactions on Knowledge and Data Engineering,, vol. 22, no. 10, pp. 1345–1359, 2010.

[25] Yoshua Bengio, "Deep learning of representations for unsupervised and transfer learning," Proc. ICML Unsupervised and Transfer Learning, pp.17-36, 2012.

[26] 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.

[27] Douglas B Paul and Janet M Baker, "The design for the Wall Street Journal-based CSR corpus," in Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, 1992, pp. 357–362

[28] 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.

[29] Tanaka, K., Kojima, H., Tomiyama, Y. and Dantsuji, M. (2001) Acoustic models of languageindependent phonetic code systems for speech processing. Spring Meeting of the Acoustical Society of Japan: Proceedings. Tokyo: Acoustical Society of Japan, 1: 191–192.