Cross-language Transfer Learning for Deep Neural Network Based Speech Enhancement

Yong Xu[1]*, Jun Du*[1]*, Li-Rong Dai*[1] *and Chin-Hui Lee*[2]

[1]National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China
[2]School of Electrical and Computer Engineering, Georgia Institute of Technology
xuyong62@mail.ustc.edu.cn, {jundu, lrdai}@ustc.edu.cn, chl@ece.gatech.edu

Abstract

In this paper, we propose a transfer learning approach to adapt a well-trained model obtained with high-resource materials of one language to another target language using a small amount of adaptation data for speech enhancement based on deep neural networks (DNNs). We investigate the performance degradation issues of enhancing noisy Mandarin speech data using DNN models already trained with only English speech materials, and vice versa. By assuming that the hidden layers of the well-trained DNN regression model as a cascade of feature extractors, we hypothesize that the first several layers should be transferable between languages. Our experimental results indicate that even with only about 1 minute of adaptation data from the resource-limited language we can achieve a considerable performance improvement over the DNN model without cross-language transfer learning.

Index Terms: speech enhancement, deep neural network, transfer learning, multi-lingual, resource-limited language

1. Introduction

Single channel speech enhancement is still a challenging task considering that the characteristics of both the speech and noise signals are very complicated in the real world environments. The traditional speech enhancement methods, such as spectral subtraction [1], Wiener filtering [2], minimum mean squared error (MMSE) estimation [3, 4] and optimally-modified log-spectral amplitude (OM-LSA) speech estimator (e.g., [5, 6]), were developed during the past several decades. However, most of them are ineffective to deal with the highly non-stationary noise (e.g., Machine Gun noise in [7]) due to the difficulty to accurately estimate the local noise spectrum [8]. The residual noise including the musical noise [9], especially at low signal-to-noise ratios (SNRs), is another problem to limit their applications in automatic speech recognition (ASR), mobile communication and hearing aids [10].

Recently, a deep learning (e.g., [11, 12]) based framework was proposed for speech enhancement and some promising results were obtained. In (e.g., [9, 13, 14]), DNN-based or stacked de-noising auto-encoder (SDA) based speech enhancement methods were proposed to map the noisy speech to the clean speech. No musical noise was found in the DNN or the SDA enhanced speech and the highly non-stationary noise could be suppressed. The DNN or SDA trained with a large collection of noise types could get much better performance than the traditional method even on unseen noise types [9, 15, 16]. Good generalization capacity of the DNN was also demonstrated in [17], where 100 noise types were used to train DNNs to classify the certain time-frequency unit to be speech-dominant or noise-dominant. In this paper, more than 100 noise types are also adopted to improve the robustness to the unseen noise types.

Nonetheless, the generalization capacity of DNN-based speech enhancement to clean speech data was not explicitly addressed in previous research. An observation is that the test performance of DNN trained with clean data of a single language is often severely degraded on a new language. Because of this mismatch, we investigate cross-language DNN-based speech enhancement in this paper. On one hand, if rich data of different languages can be collected, the multi-lingual DNN can be naturally trained with speech data of multiple languages. It should be noted that the multi-lingual DNNs here are different from the shared-hidden-layer multi-lingual DNNs (SHL-MDNNs) (e.g., [18, 19, 20]) in robust ASR, where the hidden layers are shared across many languages while the softmax layers are language dependent. In speech enhancement, the hidden layers and the linear reconstruction output layer are all shared across different languages in the multi-lingual approach. This also implies that multi-lingual DNNs trained with diversified speech data might outperform the mono-lingual DNN trained with only the language-specific data. On the other hand, in cases when a large mount of clean speech data is hard to collect for a certain language, a transfer learning strategy is proposed to adapt the well-trained DNN model from a resource-rich language to a resource-limited language. This adaptation process could be conducted by tuning the parameters of the top N layers to avoid over-fitting. Our experiments on two languages, namely English and Mandarin data, show the effectiveness of our proposed multi-lingual DNN and transfer learning approaches.

The rest of the paper is organized as follows. In Section 2 we describe the DNN-based speech enhancement system. Multi-lingual DNN learning and cross-language transfer learning are presented in Section 3. In Section 4, we present a series of experiments to assess the system performance. Finally we summarize our findings in Section 5.

2. System Overview

A block diagram of the DNN-based speech enhancement is illustrated in Fig. 1. It consists of the off-line DNN training stage and the on-line enhancement stage. In the training stage, more than 100 noise types are used to construct abundant sample pairs of the clean speech and the noisy speech. The log-power spectra features are extracted to map the noisy speech to the clean speech. The DNN training includes two steps, namely unsupervised pre-training and supervised fine-tuning. The Restricted Boltzmann Machine (RBM) based unsupervised pre-training al-

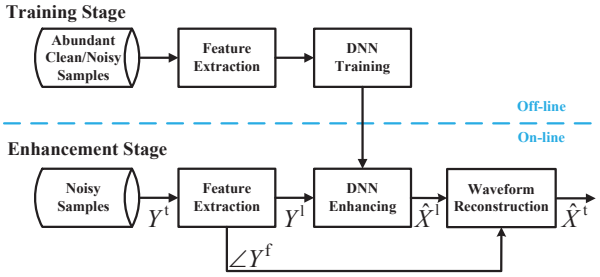

Figure 1: *A block diagram of the proposed DNN-based speech enhancement system.*

gorithm was proposed by Hinton to avoid the local optimum in the subsequent fine-tuning, especially for the DNN with many hidden layers [11]. And the conventional back-propagation algorithm is adopted to fine-tune the well initialized deep model. The MMSE criterion is designed as the object function [9]. The stochastic gradient descent is adopted for optimization in mini-batches with multiple epochs to improve learning convergence. In the enhancement stage, the noisy speech features are processed by the well-trained DNN model to predict the clean speech features. After we obtain the estimated log-power spectral features of clean speech $\hat{X}^{\mathrm{l}}$, the reconstructed spectrum $\hat{X}^{\mathrm{f}}$ could be obtained using inverse discrete Fourier transformation with the phase of input noisy speech. Finally an overlap-add method is used to synthesize the waveform of the estimated clean speech [21].

3. Cross-language DNN-based Speech Enhancement

3.1. The language mismatch problem

In [9], the cross-language performance of mandarin utterances was evaluated with an English DNN model at a specified noisy type and noise level, and it gave a preliminary cue that the language mismatch problem existed in the DNN-based speech enhancement. In this paper, we extensively investigate this problem at different SNRs under general noise environments. An English utterance example corrupted by the unseen *Exhibition* noise at SNR = 5dB was shown in the upper spectrograms in Fig. 2. Compared with the English-DNN enhanced (match testing) spectrogram shown in the upper left panel, a severe over-smoothing phenomenon was observed in the Mandarin-DNN enhanced (cross-testing) spectrogram shown in the upper middle panel. Furthermore, Fig. 2 also presents the spectrograms of a Mandarin utterance example (bottom). Compared with the Mandarin-DNN enhanced (match testing) spectrogram shown in the bottom left panel, More residue noises are left in the English-DNN enhanced (cross-testing) spectrogram shown in the bottom middle panel. These observations indicate that the language mismatch can lead to severe performance degradation. Noted that the experimental setup of the Mandarin DNN and the English DNN will be illustrated in Section 4.

To give a better understanding of those observations in Fig. 2. Fig. 3 shows the histograms of 200000 frames of the dimension-independent log-power spectra features of the clean Mandarin data and the clean English data. Obviously, the English distribution is quite different from the Mandarin distribution. Specifically, the different mean values of two distributions indicates the energy mismatch of our selected data sets.

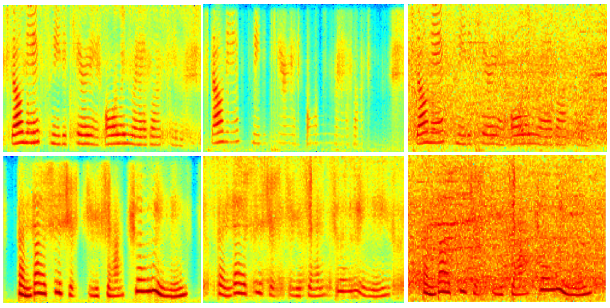

Figure 2: *Spectrograms of an English (upper) and a Mandarin (bottom) utterance example with English-DNN enhanced (upper left, PESQ=2.18), Mandarin-DNN enhanced (upper middle, PESQ=1.58), noisy (upper right, PESQ=1.58), Mandarin-DNN enhanced (bottom left, PESQ=1.70), English-DNN enhanced (bottom middle, PESQ=1.38) and noisy (bottom right, PESQ=0.96). Test on the unseen* Exhibition *noise at SNR = 5dB.*

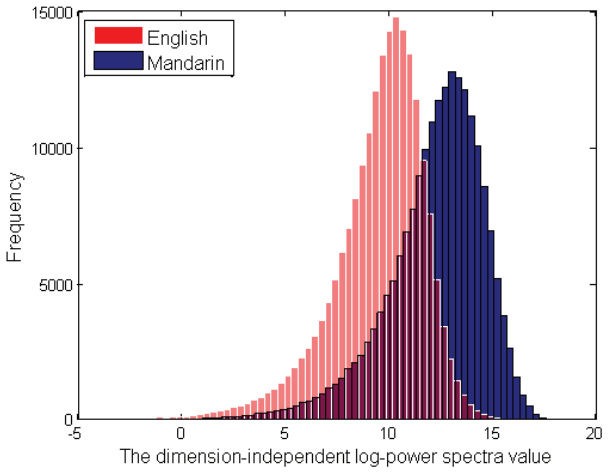

Figure 3: *Histograms of dimension-independent log-power spectra features of the clean Mandarin data and the clean English data.*

3.2. Multi-lingual DNN

Unlike the multi-lingual DNN with a language-specific softmax output layer in the ASR task (e.g., [18, 19]), all layers of the multi-lingual DNN for speech enhancement could be easily shared by different data sets without modifying the DNN structure. The key to the successful learning of the multi-lingual DNN is to train the model for all the languages simultaneously [18]. The first several hidden layers suppress the noise components by sigmoid non-linearities, while preserving the speech region [23]. After the de-noising process, the linear output layer could reconstruct the reserved speech spectrum. With the sufficient training data from each language, the performance of multi-lingual DNN could be even better than that of monolingual DNN.

3.3. Transfer learning for resource-scarce languages

Transfer learning [22] is one of the most important research topics in machine learning. It defines the ability of a system to

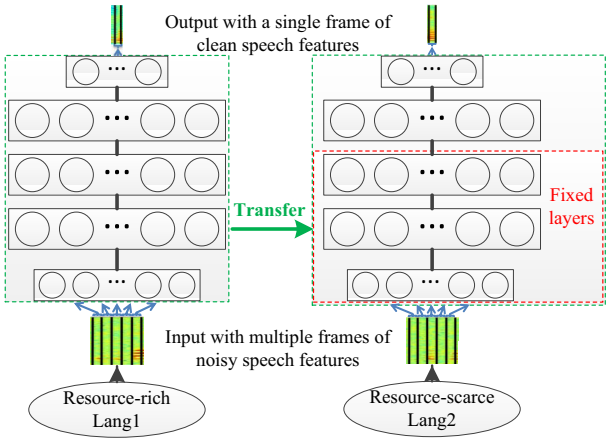

Figure 4: *Architecture of the transfer learning from the resource-rich language for the resource-scarce language.*

recognize and apply knowledge and skills learned in previous tasks to novel applications or new domains, which share some commonality. Here, to address the language mismatch issue for a resource-scarce language, the transfer learning scheme is proposed. The left DNN in Fig. 4 can be well-trained using a collection of sufficient samples of the resource-rich language, which is taken as the initialization model for the resource-limited language with a small amount of clean speech data. It should be noted that the limited clean data was fully corrupted with all noise types at various SNRs to build a multi-condition training set. With the well-initialized DNN, it can reduce the possibility to fall into the local optimum, especially when the adaptation set of the resource-scarce language is small. As the sigmoid layers play the important role to reduce noise [23], the linear output layer contains the main language-specific characteristics when reconstructing the speech spectrum, which leads to the language mismatch problem in Section 3.1. Hence, the de-noising sigmoid layers could be mostly regarded as the shared hidden layers by different languages. Only the parameters of the top layer or the top 2 layers are updated for transfer learning while keeping the other layers fixed.

4. Experimental Results and Analysis

To improve the generalization capacity of DNNs to unseen noise types, 104 noise types were used to construct the stereo data, including the four noise types, namely *AWGN*, *Babble*, *Restaurant* and *Street*, from the Aurora2 database [24], and another 100 environmental noises [25][1]. The clean English speech data was derived from the TIMIT database [26]. All 4620 utterances from the training set of the TIMIT database were corrupted with the abovementioned 104 noise types at six levels of SNR, i.e., 20dB, 15dB, 10dB, 5dB, 0dB, and -5dB, to build an English multi-condition training set, consisting of pairs of clean and noisy speech utterances. The clean Mandarin speech data is derived from our in-house corpus. The training set consists of 5400 utterances, and the length of each utterance is 3.6 seconds on average. They were used to build the corresponding Mandarin multi-condition training set. Finally 100-hour multi-condition training sets are designed for both English and Mandarin. Another 200 randomly selected utterances from the TIMIT test set and the Mandarin test set, respectively, were used to construct the test set for each combination of noise types and SNR levels. We conducted the evaluation with 3 unseen noise types[2], from the Aurora2 database [24] and the NOISEX-92 corpus [7]. An improved version of OM-LSA [5, 6], denoted as **LogMMSE**, was used for performance comparison.

All the clean speech and noise waveforms were down-sampled to 8KHz. Three objective quality measures, segmental SNR (SSNR in dB), log-spectral distortion (LSD in dB) and perceptual evaluation of speech quality (PESQ) [27], were used for evaluating the quality of the enhanced speech. Due to space limitation, we only gave selective results of those three objective measures below.

Mean and variance normalization was applied to the input and target feature vectors of the DNN. All DNN configurations were fixed at $L = 3$ hidden layers, 2048 units at each hidden layer, and 11-frame input. The number of epoch for each layer of RBM pre-training was 20. Learning rate of pre-training was 0.0005. As for the fine-tuning, learning rate was set at 0.1 for the first 10 epochs, then decreased by 10% after every epoch. Total number of epoch was 50. The mini-batch size was set to $N = 128$.

4.1. Evaluations of multi-lingual DNN

In Table 1, we compare the average PESQ results among noisy, LogMMSE, Mandarin DNN, English DNN and Multi-lingual DNN on the Mandarin or the English test set across all SNRs of the three unseen noise environments, namely *Exhibition*, *Destroyer engine* and *HF channel*. The mono-lingual DNNs are both trained on respective 100 hours data, while the multi-lingual DNN is trained on 200 hours data by combining the Mandarin data and the English data. Comparing the results of the Mandarin DNN with the English DNN on the Mandarin test set, the latter gets the worse PESQ performance degrading from 2.20 to 1.97 on average, due to the language mismatch. The mismatch problem was more observable using Mandarin DNN for cross-testing on the English test set with PESQ from 2.60 to 2.11. The average PESQ of multi-lingual DNN can be slightly better than that of corresponding mono-lingual DNNs. Furthermore, the well configured DNN-based method is superior to the LogMMSE method. Here the abundant noise types in the training set are crucial to achieve better generalization capacity to unseen noises.

4.2. Evaluations of transfer learning

In Fig. 5, transfer learning with top N layers updated is evaluated supposed that the English was the resource-limited language. It presents the average SSNR comparison on the English test set across all SNRs of the three unseen noise environments using only 72 seconds clean English data to update different parameters of DNNs with various initialization schemes. Comparing the English retrained DNN (E-DNN-retrain) with the English DNN updating all parameters initialized using the Mandarin DNN model (M-DNN-Top4), the latter gets better performance

[1]The another 100 noise types are N1-N17: Crowd noise; N18-N29: Machine noise; N30-N43: Alarm and siren; N44-N46: Traffic and car noise; N47-N55: Animal sound; N56-N69: Water sound; N70-N78: Wind; N79-N82: Bell; N83-N85: Cough; N86: Clap; N87: Snore; N88: Click; N88-N90: Laugh; N91-N92: Yawn; N93: Cry; N94: Shower; N95: Tooth brushing; N96-N97: Footsteps; N98: Door moving; N99-N100: Phone dialing.

[2]The 3 unseen environment noises for evaluation are Exhibition, Destroyer engine and HF channel. The first one noise is from the Aurora2 database and the others are collected from the NOISEX-92 corpus.

Table 1: *Average **PESQ** comparison on the **Mandarin** and **English** test set across all SNRs of the three unseen noise environments, among: Noisy, LogMMSE, Mandarin DNN (M-DNN), English DNN (E-DNN) and Multi-lingual DNN (ML-DNN).*

	Noisy	LogMMSE	E-DNN	M-DNN	ML-DNN
Mandarin	1.63	2.10	1.97	2.20	**2.22**
English	2.09	2.46	2.60	2.11	**2.61**

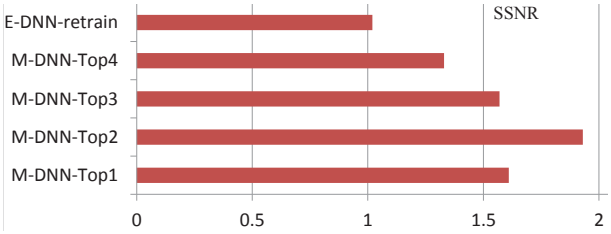

Figure 5: *Average **SSNR** comparison on the **English** test set across all SNRs of the three unseen noise environments using only 72s clean English utterances to retrain the DNN with the RBM pre-training (E-DNN-retrain) and the parameters of all or top 3/2/1 layers with the Mandarin DNN as the initialization model (denoted as M-DNN-Top4, M-DNN-Top3, M-DNN-Top2 and M-DNN-Top1, respectively.).*

because the information from parameters of Mandarin DNN can also be shared for the English language and the well initialized model can avoid over-fitting with little adaptation data. The strategy that updating the parameters of top 2 layers obtains the best performance with about 1 minute clean English data. It still can achieve a good SSNR score with only updating the parameters of the top 1 layer, indicating that the diversity between languages mostly embodied in the top layers. Furthermore, the training efficiency could also be improved with less parameters to be updated. As in Fig. 6, M-DNN-Top2 can converge fast starting from a better initialization point, but E-DNN-retrain and M-DNN-Top4 easily get stuck in over-fitting. Similar results could be found in the evaluation on the Mandarin test set. Hence, the scheme to only update the parameters of top 2 layers is adopted in the following experiments.

Table 2 presents the average LSD and PESQ comparison across all SNRs of the three unseen noise environments among transfer learning to update the top 2 layers with different clean data size of the adaptation set (0/18/72 seconds and the whole clean data in training set), transfer learning to update all layers with all clean data (denoted as All-Top4) and using all clean data to retrain (denoted as All-retrain). In the evaluation for Mandarin, the well trained English DNN was taken as the initialization model and the transfer learning with only updating the parameters of the top 2 layers was then conducted using different clean Mandarin data. After adapted with only 18s clean Mandarin data, the mismatch problem was alleviated largely with LSD from 6.90 to 5.95, and with PESQ from 1.97 to 2.09. After increasing the clean data to 72s, the LSD performance even surpassed that of All-retrain system and the PESQ performance was also comparable. With the whole clean data, transfer learning to update the parameters of all layers is just slightly better than that to update the parameters of top 2 layers, which indicates that the language-specific information mainly lies in the top layers. As for the evaluation for English, similar results could be obtained, and the final LSD and PESQ were both considerable with that of the English DNN trained on all clean data in TIMIT training set.

Table 2: *Average **LSD** and **PESQ** comparison on the **Mandarin** and **English** test set across all SNRs of the three unseen noise environments, among transfer learning to update top 2 layers with different clean data size of the adaptation set (0/18/72 seconds and the whole clean data), transfer learning to update all layers with all clean data (denoted as All-Top4) and using all clean data to retrain (denoted as All-retrain).*

	0s	18s	72s	All	All-Top4	All-retrain
LSD						
Mandarin	6.90	5.95	5.81	5.79	**5.77**	5.83
English	7.93	6.06	5.73	5.68	**5.63**	5.73
PESQ						
Mandarin	1.97	2.09	2.18	2.20	**2.21**	2.20
English	2.11	2.33	2.54	2.58	**2.60**	**2.60**

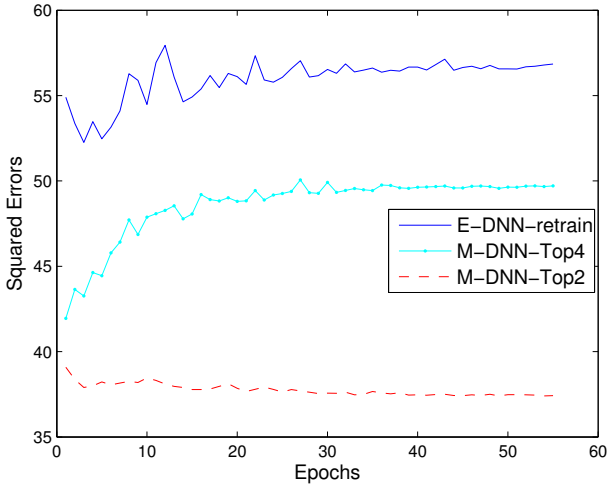

Figure 6: *Squared errors on the **English** test set using only 72s clean English utterances for transfer learning, among E-DNN-retrain, M-DNN-Top4 and M-DNN-Top2.*

5. Summary

In this paper, the language mismatch problem was analyzed and addressed for DNN-based speech enhancement. English and Mandarin databases are used for experimental design. With sufficient training samples of different languages, the multi-lingual DNN could be slightly superior to the mono-lingual DNN. However, with insufficient training samples, especially for the minority language where the clean data is difficult or expensive to collect in the real world, transfer learning was proposed to alleviate the language mismatch problem based on the sharing characteristics of DNNs between languages. We expect the proposed transfer learning approach to be applicable to addressing other mismatch conditions caused by channels, transducers and environments. Future research will be done to investigate these robustness issues in speech enhancement.

6. Acknowledgements

This work was partially supported by the National Nature Science Foundation of China (Grant No. 61273264 and No. 61305002) and the Programs for Science and Technology Development of Anhui Province (Grants No. 13Z02008-4).

7. References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustic, Speech and Signal Processing*, Vol. 27, No. 2, pp. 113-120, 1979.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," in *Proc. IEEE*, Vol. 67, No. 12, pp. 1586-1604, 1979.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 32, No.6, pp. 1109-1121, 1984.

[4] Y. Ephraim and D. Malah, "Speech enhancement using minimu mean square log spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 33, No. 2, pp. 443-445, 1985.

[5] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, Vol. 81, No. 11, pp. 2403-2418, 2001.

[6] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. on Speech and Audio Processing*, Vol. 11, No. 5, pp. 466-475, 2003.

[7] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,"

[8] K. Manohar and P. Rao, "Speech enhancement in nonstationary noise environments using noise properties," *Speech Communication*, Vol. 48, No. 1, pp. 96-109, 2006.

[9] Y. Xu, J. Du, L.-R. Dai and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, Vol. 21, No. 1, pp. 65-68, 2014.

[10] J. Benesty, S. Makino, and J. D. Chen, *Speech Enhancement*, Springer, 2005.

[11] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, Vol. 313, No. 5786, pp. 504-507, 2006.

[12] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, Vol. 2, No. 1, pp. 1-127, 2009.

[13] B.-Y. Xia and C.-C. Bao, "Speech enhancement with weighted denoising Auto-Encoder," *Proc. Interspeech*, pp. 3444-3448, 2013.

[14] X.-G. Lu and Y. Tsao and S. Matsuda and C. Hori, "Speech enhancement based on deep denoising Auto-Encoder," *Proc. Interspeech*, pp. 436-440, 2013.

[15] B.-Y. Xia and C.-C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Communication*, Vol. 60, pp. 13-29, 2014.

[16] Y. Xu, J. Du, L.-R. Dai and C.-H. Lee, "Global variance equalization for improving deep neural network based speech enhancement," *Proc. ChinaSIP*, 2014.

[17] Y. X. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 21, No. 7, pp. 1381-1390, 2013.

[18] J. T. Huang, J. Li, D. Yu, L. Deng and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," *Proc. ICASSP*, pp. 7304-7308, 2013.

[19] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin and J. Dean, "Multilingual acoustic models using distributed deep neural networks," *Proc. ICASSP*, pp. 8619-8623, 2013.

[20] A. Ghoshal, P. Swietojanski and S. Renals, "Multilingual training of deep neural networks," *Proc. ICASSP*, pp. 7319-7323, 2013.

[21] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions," *Proc. Interspeech*, pp. 569-572, 2008.

[22] S. J. Pan and Q. Yang, "A survey on transfer learning. Knowledge and Data Engineering," *IEEE Trans. on Knowledge and Data Engineering*, Vol. 22, No. 10, pp. 1345-1359, 2010.

[23] S. I. Tamura, "An analysis of a noise reduction neural network," *Proc. ICASSP*, pp. 2001-2004, 1989.

[24] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the preformance evaluations of speech recognition systems under noisy conditions," *Proc. ISCA ITRW ASR*, pp. 181-188, 2000.

[25] G. Hu, 100 nonspeech environmental sounds, 2004. http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html.

[26] J. S. Garofolo, *Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database*, NIST Tech Report, 1988.

[27] ITU-T, Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *International Telecommunication Union-Telecommunication Standardisation Sector*, 2001.