SPEECH SEPARATION BASED ON SIGNAL-NOISE-DEPENDENT DEEP NEURAL NETWORKS FOR ROBUST SPEECH RECOGNITION

Yan-Hui Tu[1], *Jun Du*[1], *Li-Rong Dai*[1], *Chin-Hui Lee*[2]

[1]University of Science and Technology of China, HeFei, AnHui, P.R. China
[2]Georgia Institute of Technology, Atlanta, Georgia, USA

tuyanhui@mail.ustc.edu.cn, {jundu,lrdai}@ustc.edu.cn, chl@ece.gatech.edu

ABSTRACT

In this paper, we propose a new signal-noise-dependent (SND) deep neural network (DNN) framework to further improve the separation and recognition performance of the recently developed technique for general DNN-based speech separation. We adopt a divide and conquer strategy to design the proposed SND-DNNs with higher resolutions that a single general DNN could not well accommodate for all the speaker mixing variabilities at different levels of signal-to-noise ratios (SNRs). In this study two kinds of SNR-dependent DNNs, namely positive and negative DNNs, are trained to cover the mixed speech signals with positive and negative SNR levels, respectively. At the separation stage, a first-pass separation using a general DNN can give an accurate SNR estimation for a model selection. Experimental results on the Speech Separation Challenge (SSC) task show that SND-DNNs could yield significant performance improvements for both speech separation and recognition over a general DNN. Furthermore, this purely front-end processing method achieves a relative word error rate reduction of 11.6% over a state-of-the-art recognition system where a complicated joint decoding framework needs to be implemented in the back-end.

Index Terms— single-channel speech separation, robust speech recognition, deep neural networks, semi-supervised mode

1. INTRODUCTION

Speech separation aims at separating the voice of each speaker when multiple speakers talk simultaneously. It is important for many applications, for example automatic speech recognition (ASR). While significant progress has been made in improving the noise robustness of ASR systems, most techniques focus on improving the performance of the back-end recogniser. In this study, we use the separation system as our front-end pre-processor for ASR. So the performance of the ASR system depends heavily on the quality of acoustic pre-processing. The separating algorithms can be often classified into unsupervised and supervised modes. In the former, speaker identities and the reference speech of each speaker are not available in the training stage, while the information of both the target and the interfering speakers is provided in the supervised modes.

One broad class of single-channel speech separation is the so-called computational auditory scene analysis (CASA) [1], usually in an unsupervised mode. CASA-based approaches [2]-[6], use the psychoacoustic cues, such as pitch, voice onset/offset, temporal continuity, harmonic structures, and modulation correlation, to segregate a voice of interest by masking the interfering sources. For example, in [5], pitch and amplitude modulation were adopted to separate the voiced portions of co-channel speech. In [6], unsupervised clustering was used to separate speech regions into two speaker groups by maximizing the ratio of between-cluster and within-cluster distances. Recently, a data-driven approach [7] separates the underlying clean speech segments by matching each mixed speech segment against a composite training segment.

In the supervised approaches, speech separation is often formulated as an estimation problem based on:

$$\boldsymbol{x}^{\mathrm{m}} = \boldsymbol{x}^{\mathrm{t}} + \boldsymbol{x}^{\mathrm{i}} \tag{1}$$

where $\boldsymbol{x}^{\mathrm{m}}$, $\boldsymbol{x}^{\mathrm{t}}$, $\boldsymbol{x}^{\mathrm{i}}$ are speech signals of the mixture, target speaker, and interfering speaker, respectively. To solve this under-determined equation, a general strategy is to represent the speakers by two models, and use a certain criterion to reconstruct the sources given the single mixture. An early study in [8] adopted a factorial hidden Markov model (FHMM) to describe a speaker, and the estimated sources are used to generate a binary mask. To further impose temporal constraints on speech signals for separation, the work in [9] investigates the phone-level dynamics using HMMs [10]. For FHMM-based speech separation, 2-D Viterbi algorithms and approximations have been used to estimate the inference [11]. In [12], FHMM was adopted to model vocal tract characteristics for detecting pitch to reconstruct speech sources. In [13, 14, 15] Gaussian mixture models (GMMs) were employed to model speakers, and the minimum mean squared error (MMSE) or maximum a *posteriori* (MAP) estimator is used to recover the speech signals. The factorial-max vector quantization model (MAXVQ) was also used to infer the mask signals in [16]. Other popular approaches include nonnegative matrix factorization (NMF) based models [17].

Recently, speech separation based on deep learning approaches becomes increasingly popular, which can be divided into two broad classes. One is in a supervised mode, where deep neural networks (DNNs) or recurrent neural networks (RNNs) [18] are adopted to separate the mixed speech given the information of the target speaker, interfering speaker, and even the signal-to-noise ratio (SNR). The other one is in a semi-supervised mode where only the information of the target speaker is provided. Our recent work [19, 20, 21] belongs to the latter. In [19, 20], we solve the separation problem in Eq. (1) by using DNN to directly model the highly nonlinear relationship among speech features of the target speaker, the interference speaker and the mixed signals. Its effectiveness has also been verified for robust speech recognition [21]. As our DNN approach is semi-supervised, a large mount of training data with different interfering speakers at different SNRs can be included to address the problem of unseen information. However a single general DNN might not

This work was supported by the National Natural Science Foundation of China under Grants No. 61305002.

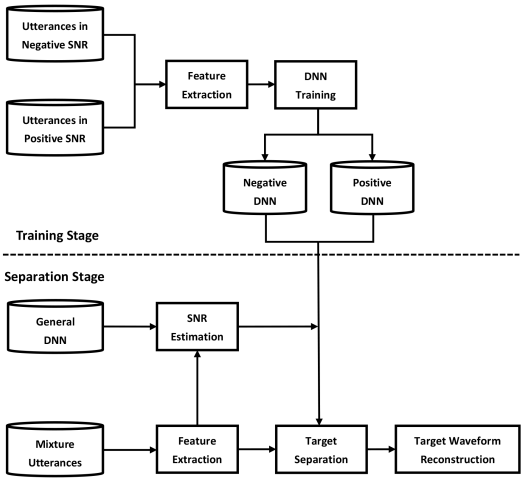

Fig. 1. Development flow for speech separation system.

accommodate all the variabilities well. In this study we adopt a divide and conquer strategy to design signal-noise-dependent DNNs (SND-DNNs) with a detailed resolutions. Two SND-DNNs, namely positive and negative DNNs, are trained to cover the mixed speech with positive SNRs and negative SNRs, respectively. At the separation stage, the first-pass separation using a general DNN can give an accurate SNR estimation for the follow-up model selection.

The evaluation results on the Speech Separation Challenge (SSC) corpus [22] show that the proposed SND-DNNs approach significantly outperforms the general DNN approach [21] in terms of both separation and recognition performance. Furthermore, our purely front-end only pre-processing method achieves significant performance improvements over the state-of-the-art IBM system [23, 24] and a comparable performance with recent work in [25], where a complicated joint decoding framework or/and DNN based acoustic modeling should be implemented in the back-end.

The rest of the paper is organized as follows. In Section 2, we give a system overview. In Section 3, we propose SND-DNN based speech separation. In Section 4, we report experimental results. Finally we conclude our findings in Section 5.

2. SYSTEM OVERVIEW

In this section, both the speech separation and the ASR systems are introduced. First, an overall flowchart of our proposed speech separation system is illustrated in Fig. 1. In the training stage, the DNN as a regression model is trained by using log-power spectra features from pairs of mixed signal and the sources. Two SND-DNNs, namely positive DNN and negative DNN, are trained using mixture utterances with positive SNRs and negative SNRs, respectively. In the separating stage, we use a general DNN to perform first-pass separation for SNR estimation of the mixture. Then based on the estimated SNR, the positive or negative DNN is selected for the second-pass separation. Meanwhile, in Fig. 2, the development flow of the speech recognition system is given. In the training stage, the acoustic model using Gaussian mixture continuous density HMMs (denoted as GMM-HMMs) is trained from the clean speech of the target speaker using mel-frequency cepstral coefficients (MFCCs) under the maximum likelihood (ML) criterion. In the recognition stage, the mixture

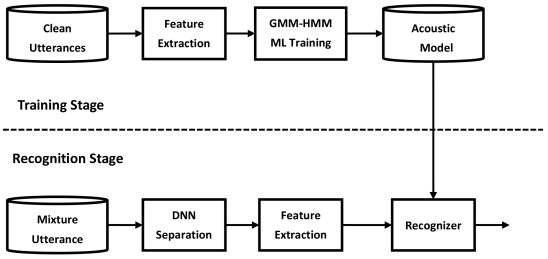

Fig. 2. Development flow for speech recognition system.

utterance is first preprocessed by speech separation based on SND-DNNs to extract the speech waveforms of the target speaker. Then conventional feature extraction and recognition follow. In the next section, the detail of SND-DNNs is elaborated.

3. SPEECH SEPARATION BASED ON SND-DNNS

As the procedures for training positive DNN or negative DNN are the same, we first introduce training of the general DNN which separates the mixture utterances in all SNRs, and then SNR estimation based on separation results.

3.1. DNN for predicting the target and interference

In [20], DNN was adopted as a regression model to predict the log-power spectra features of the target and interference speakers given the input log-power spectra features of mixed speech with acoustic context as shown in Fig. 3. These spectra features provide perceptually relevant parameters. The acoustic context information along both time axis (with multiple neighboring frames) and frequency axis (with full frequency bins) can be fully utilized by DNN to improve the continuity of the estimated clean speech while the conventional GMM-based approach does not effectively model the temporal dynamics of speech. As training of this regression DNN requires a large amount of time-synchronized stereo-data with target and mixed speech pairs, the mixed speech utterances are synthesized by corrupting the clean speech utterances of the target speaker with interferers at different SNR levels (here we consider interfering speech as noise) based on Eq. (1).

Training of DNN consists of unsupervised pre-training and supervised fine-tuning. Pre-training treats each consecutive pair of layers as a restricted Boatsman machine (RBM) [26] while the parameters of RBMs are trained layer by layer with an approximate contrastive divergence algorithm [27]. For supervised fine-tuning, we aim at jointly minimizing the mean squared error between the DNN output and the reference clean features of both the target and interference speakers:

$$E = \frac{1}{N}\sum_{n=1}^{N}(\|\hat{\boldsymbol{x}}_n^{\mathrm{t}} - \boldsymbol{x}_n^{\mathrm{t}}\|_2^2 + \|\hat{\boldsymbol{x}}_n^{\mathrm{i}} - \boldsymbol{x}_n^{\mathrm{i}}\|_2^2) \quad (2)$$

where $\hat{\boldsymbol{x}}_n^{\mathrm{t}}$ and $\boldsymbol{x}_n^{\mathrm{t}}$ are the n^{th} D-dimensional vectors of estimated and reference clean features of the target speaker, respectively, while $\hat{\boldsymbol{x}}_n^{\mathrm{i}}$ and $\boldsymbol{x}_n^{\mathrm{i}}$ are the corresponding versions for interference.

In the conventional supervised approaches for speech separation, e.g., GMM-based method [15], both the target and interference in the

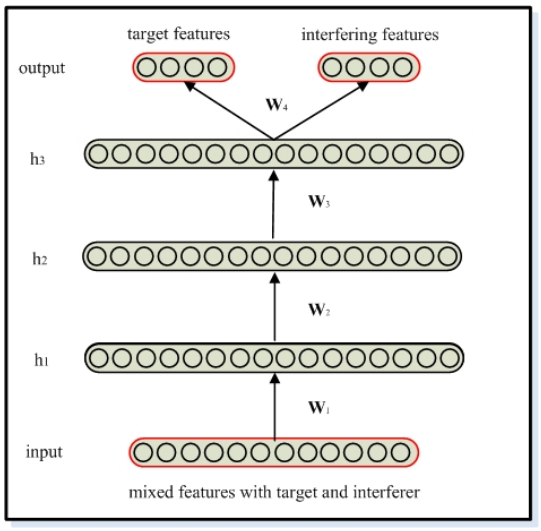

Fig. 3. DNN architecture.

separation stage should be well modeled by GMMs with the corresponding speech data in the training stage. In this paper, we mainly focus on speech separation of a target speaker in a *semi-supervised* mode, where the interferer in the separation stage is assumed unknown in the training stage. Obviously, GMM cannot be easily applied here. On the other hand for the DNN-based approach, multiple interfering speakers mixed with a target speaker in the training stage can well predict unseen interferers in the separation stage [19].

3.2. SNR estimation

The separated target and interference utterances by the general DNN can be used for SNR estimation of the current utterance according to the following equation:

$$\text{SNR} = 10\log\left(\frac{\sum_m x_{\text{t}}^2[m]}{\sum_m x_{\text{i}}^2[m]}\right) \tag{3}$$

where $x_{\text{t}}[m]$ and $x_{\text{i}}[m]$ are the m^{th} samples of reconstructed target and interference signals in time domain, respectively. With this estimated SNR, the corresponding SND-DNN can be selected for the second-pass speech separation. In this work, we simply set 0 dB as a threshold to select positive DNN or negative DNN. Using only two SND-DNNs could guarantee both high model resolution and accurate model selection.

4. EXPERIMENTS

Experiments were conducted on the SSC (Speech Separation Challenge) corpus [22] for recognizing a few keywords from simple *target* sentences when presented with a simultaneous *masker* sentence with a very similar structure [23]. All the training and test materials were drawn from the GRID corpus [28]. There were 34 speakers for both training and test, including 18 males and 16 females. For the training set, 500 utterances were randomly selected from the GRID corpus for each speaker. The test set of the SSC corpus consists of two-speaker mixtures at a range of target-to-masker ratios (TMRs)

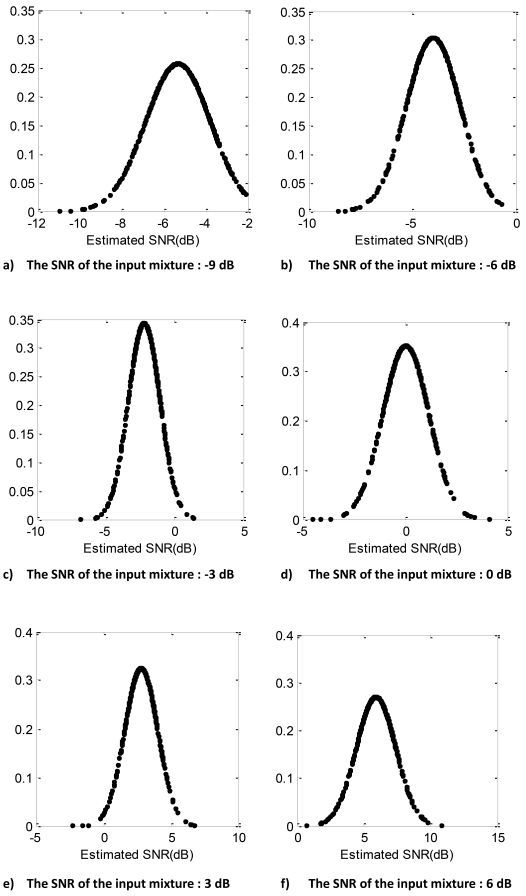

Fig. 4. Distribution of estimated SNR of input mixtures with different SNR levels.

from -9dB to 6dB with an increment of 3dB. For training the general DNN of each target speaker, all the utterances of the target speaker in the training set were used while the corresponding mixtures were generated by adding randomly selected interferers to the target speech at SNRs ranging from -10 dB to 10 dB with an increment of 1 dB. The mixture speech data with SNRs ranging from -10 dB to 0 dB were used to train the negative DNN while the positive DNN were trained using the mixture speech with SNRs ranging from 0 dB to 10 dB.

As for signal analysis, all waveforms were down-sampled from 25kHz to 16kHz, and the frame length was set to 512 samples (or 32 msec) with a frame shift of 256 samples. A short-time Fourier transform was used to compute the discrete Fourier transform (DFT) of each overlapping windowed frame. Then 257-dimensional log-power spectra features were used to train DNNs. The separation performance was evaluated using a short-time objective intelligibility (STOI) [29] and the recognition accuracy. The DNN architecture used in all experiments was 1799-2048-2048-2048-514, which denoted that the sizes were 1799 (257*7) for the input layer, 2048 for three hidden layers, and 514(257*2) for the output layer. The number of epoch for each layer of RBM pre-training was 20 while the learning rate of pre-training was 0.0005. For fine-tuning, the learn-

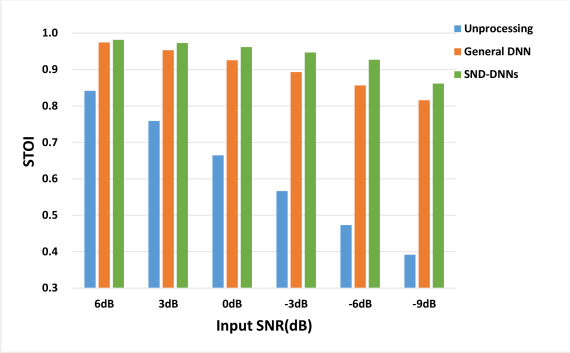

Fig. 5. Separation performance (STOI) comparison of different approaches averaged across all 34 testing target speakers.

ing rate was set at 0.1 for the first 10 epochs, then decreased by 10% after every epoch. The total number of epoch was 50 and the mini-batch size was set to 128. Input features of DNNs were globally normalized to zero mean and unit variance. Other parameter settings can be found in [30].

As for the recognition system, the feature vector consists of 39-dimensional MFCCs, i.e., 12 mel-cepstral coefficients and the logarithmic energy plus the corresponding first and second order time derivatives. Each word was modeled by a whole-word left-to-right HMMs with 32 Gaussian mixtures per state. The chosen number of states for each word can be referred to [23].

4.1. Experiments on SNR estimation

The separation performance using SND-DNNs depends highly on how accurate the SNR estimation of the mixture utterance is. Fig. 4 shows the distributions of estimated SNRs of the test data with different input SNRs. Several observations can be made. First, for all the testing cases except the input SNR of 0 dB, our SNR estimation based on the separation results of the general DNN could give accurate decisions on positive SNR or negative SNR. As for the 0 dB cases, there was no significant influence in the final decision because 0 dB training data were included for both positive DNN and negative DNN. Second, all distributions were unimodal. When the input SNR was above -3 dB, the distribution was centered exactly on the input SNR which indicated that a good estimation was given by our approach. But for input SNR below -3 dB, e.g., Figures 4(a) and 4(b), the separation performance was degraded which led to the center shift and a large variance of the distribution. Overall, our proposed SNR estimation approach was accurate enough to make the subsequent decision for the two SND-DNNs.

4.2. Results on speech separation

Fig. 5 lists a STOI comparison of different approaches averaged across all 34 target speakers on the test set. The number of interfering speakers in the training stage was set to 10, which resulted in about 100 hours of mixed speech for each target speaker. A total of 34 general DNNs and 68 SND-DNNs were trained for all target speakers. Based on those results, the general DNN approach yielded a very significant improvements of STOI performance over the unprocessed input mixtures. Meanwhile, our proposed SND-DNNs approach consistently outperformed the general DNN approach especially for low SNR cases. For example at SNR = -3 dB, the STOI was improved from 0.89 to 0.95.

Table 1. The performance (word accuracy in %) comparison of the baseline, the general DNN, the SND-DNNs approach, and IBM systems averaged across the mixture data of the test set.

	6dB	3dB	0dB	-3dB	-6dB	-9dB	Avg.
16KHz waveform							
Baseline	49.1	34.2	22.9	13.7	10.2	8.0	23.0
DNN	92.6	89.7	86.7	81.3	75.1	69.9	82.6
SND-DNNs	93.1	90.9	89.3	87.6	84.7	75.9	86.9
25KHz waveform							
Baseline	63.3	47.5	35.2	24.0	17.0	12.0	33.2
SND-DNNs	**94.9**	**93.6**	**92.4**	**90.6**	**87.0**	**81.9**	**90.1**
IBM	93.0	92.5	91.5	89.5	87.0	79.0	88.8

4.3. Results on robust speech recognition

Finally, the effectiveness of the SND-DNNs based separation approach is further verified for robust speech recognition. In Table 1, we report the performance (word accuracy in %) comparison of the baseline, the general DNN, the SND-DNNs, and IBM systems averaged across the mixture data of the test set. As our experiments in this study and previous work [19, 20, 21] are mainly conducted on 16KHz waveform, to perform a fair comparison with the IBM results [24] on 25KHz waveform, we give our SND-DNNs results for both 16KHz and 25KHz waveforms.

The general DNN achieved significant performance improvements over the baseline system without speech separation. On top of the general DNN, SND-DNNs yielded consistently additional performance gains for all testing cases, especially at low SNRs, e.g., at -3 dB, a relative word error rate (WER) reduction of 38.6% was observed. In average, an absolute 4.3% WER was reduced. Furthermore, our SND-DNNs approach consistently outperformed the IBM system under all SNRs. For example, relative WER reductions of 27.1% and 13.8% were yielded at 6 dB and -9 dB, respectively. And overall a relative WER reduction of 11.6% averaged across the whole test set was achieved. By considering that the IBM system used both speech separation in the front-end and a complicated joint decoding framework in the back-end, our purely front-end approach based on SND-DNNs is quite effective and we expect additional post-processing could further increase the word accuracy.

5. CONCLUSION AND FUTURE WORK

We have proposed signal-noise-dependent DNNs to achieve high model resolutions. As a specific implementation, two SND-DNNs, namely positive and negative DNNs, demonstrate that the proposed SND-DNNs approach could be more effective than the general DNN approach on speech separation and robust speech recognition for all testing cases. Furthermore, our purely front-end processing method is easier to implement and achieves a better recognition performance than the state-of-the-art IBM super-human system where a complicated joint decoding framework needs to be implemented in the back-end. Our future work includes further improving the separation performance at low SNRs by using more detailed SND-DNNs and even gender-dependent DNNs, and also adopting deep learning approaches for the back-end of the ASR system.

6. REFERENCES

[1] D. L. Wang and G. J. Brown, *Computational, Auditory Scene Analysis: Principles,Algorithms and Applications*, Wiley-IEEE Press, Hoboken, 2006.

[2] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Netw.*, Vol. 10, No. 3, pp. 684-697, 1999.

[3] M. Wu, D. L. Wang, and G. J. Brown, "A multi-pitch tracking algorithm for noisy speech," *IEEE Trans. Audio Speech Processing*, Vol. 11, No. 3, pp. 229-241, 2003.

[4] Y. Shao and D. L. Wang, "Model-based sequential organization in cochannel speech," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 14, No. 1, pp. 289-298, 2006.

[5] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 18, No. 8, pp. 2067-2079, 2010.

[6] K. Hu and D. L. Wang, "An unsupervised approach to cochannel speech separation," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 21, No. 1, pp. 120-129, 2013.

[7] J. Ming, R. Srinivasan, D. Crookes, and A. Jafari, "CLOSE–a data-driven approach to speech separation," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 21, No. 7, pp. 1355-1368, 2013.

[8] S. Roweis, "One microphone source separation," *Adv. Neural Inf. Process. Syst. 13*, 2000, pp. 793-799.

[9] R. Weiss and D. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Comput. Speech Lang.*, Vol. 24, pp. 16-29, 2010.

[10] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257C286, 1989.

[11] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Single-channel multitalker speech recognition," *IEEE Signal Process. Mag.*, Vol. 27, No. 6, pp. 66-80, 2010.

[12] M. Stark, M. Wohlmayr, and F. Pernkopf, "Source-filter-based single-channel speech separation using pitch information," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 19, No. 2, pp. 242-255, 2011.

[13] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 15, No. 6, pp. 1766-1776, 2007.

[14] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft masking filtering," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 15, No. 8, pp. 2299-2310, 2007.

[15] K. Hu and D. L. Wang, "An iterative model-based approach to cochannel speech separation", *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 14, 2013.

[16] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monoaural speech separation based on MAXVQ and CASA for robust speech recognition," *Computer Speech and Language*, Vol. 24, pp. 30-44, 2010.

[17] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization factorization," *Proc. INTERSPEECH*, 2006, pp. 2614-2617.

[18] P.-S. Huang, M. Kim, M. H. Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," *Proc. ICASSP*, 2014, pp. 1581-1585.

[19] J. Du, Y.-H. Tu, Y. Xu, L.-R. Dai, and C.-H. Lee, "Speech separation of a target speaker based on deep neural networks," *Accepted by Proc. ICSP*, 2014.

[20] Y.-H. Tu, J. Du, Y. Xu, L.-R. Dai, and C.-H. Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," *Accepted Proc. ISCSLP*, 2014.

[21] Y.-H. Tu, J. Du, Y. Xu, L.-R. Dai, and C.-H. Lee, "Deep neural network based speech separation for robust speech recognition," *Accepted by Proc. ICSP*, 2014.

[22] M. Cooke and T.-W. Lee, Speech Separation Challenge, 2006. [http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge.htm]

[23] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech and Language*, Vol. 24, No. 1, pp. 1-15, 2010.

[24] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: a graphical modeling approach," *Computer Speech and Language*, Vol. 24, No. 1, pp. 44-66, 2010.

[25] C. Weng, D. Yu, M. Seltzer, and J. Droppo, "Single-channel mixed speech recognition using deep neural networks," *Proc. ICASSP*, 2014, pp. 5669-5673.

[26] Y. Bengio, "Learning deep architectures for AI," *Foundat. and Trends Mach. Learn.*, Vol. 2, No. 1, pp. 1-127, 2009.

[27] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, Vol. 18, pp. 1527-1554, 2006.

[28] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, Vol. 120, No. 5, pp. 2421-2424, 2006.

[29] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," *Proc. ICASSP*, 2010, pp. 4214-4217.

[30] G. Hinton, "A practical guide to training restricted Boltzmann machines," UTML TR 2010-003, University of Toronto, 2010.