Speech Separation Based on Improved Deep Neural Networks with Dual Outputs of Speech Features for Both Target and Interfering Speakers

Yanhui Tu[1], *Jun Du*[1], *Yong Xu*[1], *Lirong Dai*[1], *Chin-Hui Lee*[2]

[1]University of Science and Technology of China
[2]Georgia Institute of Technology

{tuyanhui,xuyong62}@mail.ustc.edu.cn, {jundu,lrdai}@ustc.edu.cn, chl@ece.gatech.edu

Abstract

In this paper, a novel deep neural network (DNN) architecture is proposed to generate the speech features of both the target speaker and interferer for speech separation. DNN is adopted here to directly model the highly nonlinear relationship between speech features of the mixed signals and the two competing speakers. With the modified output speech features for learning the parameters of the DNN, the generalization capacity to unseen interferers is improved for separating the target speech. Meanwhile, without any prior information from the interferer, the interfering speech can also be separated. Experimental results show that the proposed new DNN enhances the separation performance in terms of different objective measures under the semi-supervised mode where the training data of the target speaker is provided while the unseen interferer in the separation stage is predicted by using multiple interfering speakers mixed with the target speaker in the training stage.

Index Terms: single-channel speech separation, 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 such as speech communication and automatic speech recognition. In this study, we focus on the separation of two voices from a single mixture, namely single-channel (or co-channel) speech separation. Based on the information used the algorithms can be classified into two categories: unsupervised and supervised modes. In the former, speaker identities and the reference speech of each speaker for pre-training are not available, while the information of both the target and 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, onset/offset, temporal continuity, harmonic structures, and modulation correlation, and segregate a voice of interest by masking the interfering sources. For example, in [5], pitch and amplitude modulation are adopted to separate the voiced portions of cochannel speech. In [6], unsupervised clustering is used to separate speech regions into two speaker groups by maximizing the ratio of between-cluster distance and within-cluster distance. 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 underdetermined 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] adopts 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. For FHMM based speech separation, 2-D Viterbi algorithms and approximations have been used to perform the inference [10]. In [11], FHMM is adopted to model vocal tract characteristics for detecting pitch to reconstruct speech sources. In [12, 13, 14] Gaussian mixture models (GMMs) are 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) is also used to infer the mask signals in [15]. Other popular approaches include nonnegative matrix factorization (NMF) based model [16].

One recent work [17] uses deep neural networks (DNNs) to solve the separation problem in Eq. (1) in an alternative way. DNN is adopted to directly model the highly non-linear relationship between speech features of a target speaker and the mixed signals. Eq. (1) plays the role of synthesizing a large amount of the mixed speech for DNN training, given the speech sources of the target speaker and interfering speaker. This framework avoids the difficult relationship based on Eq. (1) using complex models for both the target and interfering speakers and significantly outperforms the GMM-based separation in [14] due to the powerful modeling capability of DNN. In this paper, we propose a novel architecture of DNN which is designed to predict the speech features of both the target speaker and interferer. With this newly defined objective function aiming at minimizing the mean squared error between the DNN output and the reference clean features of both the target and the interfering speakers. It leads to an improved generalization capacity to unseen interferers for separating the target speech signal. Meanwhile, without any information from the interferer, the interference speech can also be well separated for developing new algorithms and applications.

The remainder of the paper is organized as follows. In Section 2, we give a system overview. In Section 3, we introduce the details of DNN-based speech separation. In Section 4, we report experimental results and finally we conclude the paper in Section 5.

978-1-4799-4219-0/14/$31.00

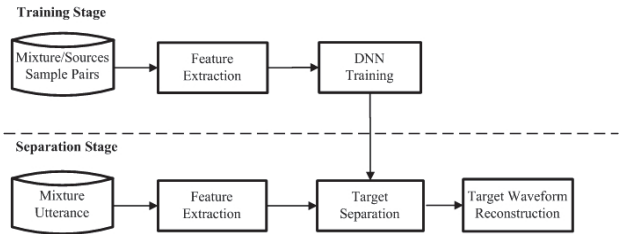

Figure 1: Overall development flow and architecture.

2. System Overview

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 spectral features from pairs of mixed signal and the sources. Note that in this work there are only two speakers in the mixed signal, namely the target speaker and the interfering speaker. In the separation stage, the log-power spectral features of the mixture utterance are processed by the well-trained DNN model to predict the speech feature of the target speaker. Then the reconstructed spectra could be obtained using the estimated log-power spectra from DNN and the original phase of mixed speech. Finally, an overlap add method is used to synthesize the waveform of the estimated target speech [18]. In the next section, the details of two types of DNN architectures are elaborated.

3. DNN-based Speech Separation

3.1. DNN-1 for predicting the target

In [17], DNN is adopted as a regression model to predict the log-power spectral features of the target speaker given the input log-power spectral features of mixed speech with acoustic context as shown in Fig. 2. These spectral 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 estimated clean speech while the conventional GMM-based approach do not model the temporal dynamics of speech. As the 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 signal-to-noise (SNR) levels (here we consider interfering speech as noise) based on Eq. (1). Note that the generalization to different SNR levels in the separation stage can be well addressed by the full coverage of SNR levels in the training stage inherently.

Training of DNN consists of unsupervised pre-training and supervised fine-tuning. The pre-training treats each consecutive pair of layers as a restricted Boltzmann machine (RBM) [20] while the parameters of RBM are trained layer by layer with the approximate contrastive divergence algorithm [19]. For the supervised fine-tuning, we aim at minimizing mean squared error between the DNN output and the reference clean features of the target speaker:

$$E_1 = \frac{1}{N}\sum_{n=1}^{N} \|\hat{\boldsymbol{x}}_n^{\mathrm{t}}(\boldsymbol{x}_{n\pm\tau}^{\mathrm{m}}, \boldsymbol{W}, \boldsymbol{b}) - \boldsymbol{x}_n^{\mathrm{t}}\|_2^2 \qquad (2)$$

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. $\boldsymbol{x}_{n\pm\tau}^{\mathrm{m}}$ is a $D(2\tau+1)$-dimensional vector of input mixed features with neighbouring left and right τ frames as the acoustic context. $\boldsymbol{W}$ and $\boldsymbol{b}$ denote all the weight and bias parameters. The objective function is optimized using back-propagation with a stochastic gradient descent method in mini-batch mode of N sample frames. As this DNN only predicts the target speech features in the output layer, we denote it as **DNN-1**.

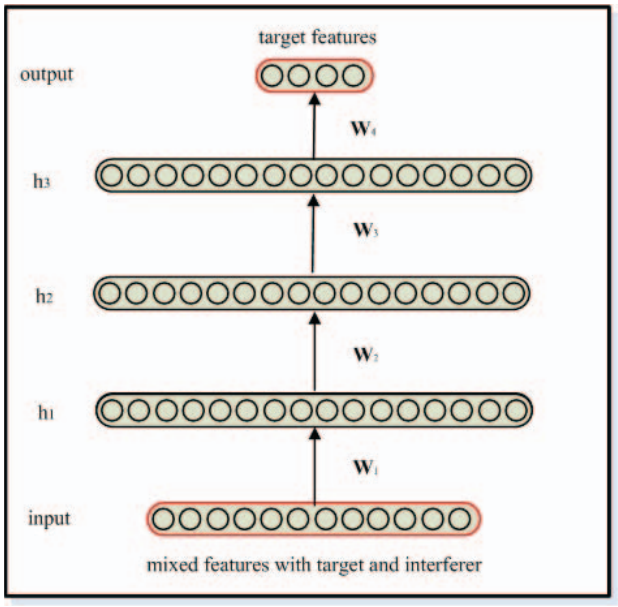

Figure 2: DNN-1 architecture.

3.2. DNN-2 for predicting both the target and interference

In this work, we design a new DNN architecture for speech separation which is illustrated in Fig. 3. The main difference from Fig. 2 is that the new DNN can predict both the target and interference in the output layer which is denoted as **DNN-2**. The pre-training of DNN-2 is exactly the same as that of DNN-1 while the supervised fine-tuning is conducted by jointly minimizing the mean squared error between the DNN output and the reference clean features of both the target and interference:

$$E_2 = \frac{1}{N}\sum_{n=1}^{N}(\|\hat{\boldsymbol{x}}_n^{\mathrm{t}} - \boldsymbol{x}_n^{\mathrm{t}}\|_2^2 + \|\hat{\boldsymbol{x}}_n^{\mathrm{i}} - \boldsymbol{x}_n^{\mathrm{i}}\|_2^2) \qquad (3)$$

where $\hat{\boldsymbol{x}}_n^{\mathrm{i}}$ and $\boldsymbol{x}_n^{\mathrm{i}}$ are the n^{th} D-dimensional vectors of estimated and reference clean features of the interference, respectively. The second term of Eq. (3) can be considered as a regularization term for Eq. (2), which leads to better generalization capacity for separating the target speaker. Another benefit from DNN-2 is the inference can also be separated as a by-product for developing new algorithms and other applications.

3.3. Semi-supervised mode

In the conventional supervised approaches for speech separation, e.g., GMM-based method [14], both the target and interference in the separation stage should be well modeled by GMM with the corresponding speech data in the training stage. In [17], it is already demonstrated that DNN-1 can achieve consistent and significant improvements over the GMM-based approach in the supervised mode. In this paper, we only focus

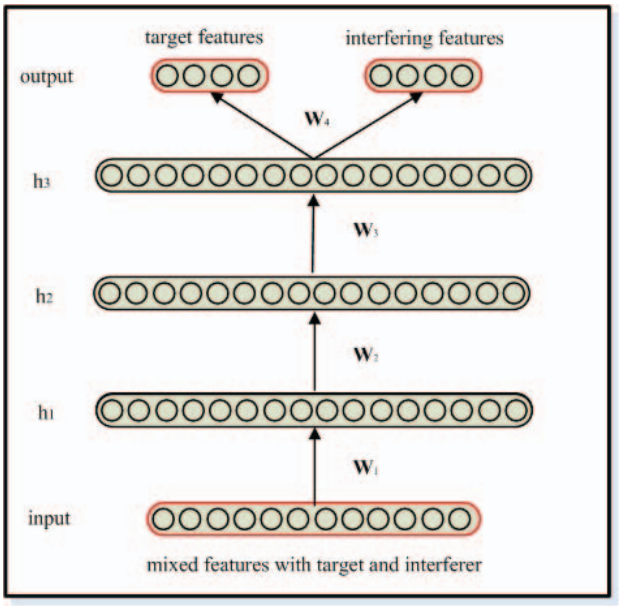

Figure 3: DNN-2 architecture.

on speech separation of a target speaker in a *semi-supervised* mode for both DNN-1 and DNN-2, where the interferer in the separation stage is excluded 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 an unseen interferer in the separation stage [17].

4. Experiments

Our experiments were conducted on our in-house Mandarin corpus. We have two target speakers, namely one male and one female. For each target speaker, 200 utterances were used for training with 30 utterances for testing. The interfering speakers were randomly selected from a large set with thousands of speakers. For training of DNNs, all the utterances of the target speakers in the training set were used while the corresponding mixtures were generated by adding randomly selected interferers to the target speech at SNR levels ranging from -15 dB to 15 dB with an increment of 5 dB. The test set for each target speaker consisted of 25 male and 25 female interferers, which are not included in the training stage. Then the mixtures are generated by the target speaker and each interferer at SNRs from -9 dB to 6 dB with an increment of 3 dB for evaluation.

As for signal analysis, the sampling rate of speech waveforms was 16kHz, and the frame length was set to 512 samples (or 32 msec) with a frame shift of 256 samples. A short-time Fourier analysis was used to compute the DFT of each overlapping windowed frame. Then 257-dimensional log-power spectra features were used to train DNNs. The separation performance was evaluated using three measures, namely output SNR [14], a short-time objective intelligibility (STOI) [22], and perceptual evaluation of speech quality (PESQ) [23]. STOI is shown to be highly correlated to human speech intelligibility while PESQ has a high correlation with subjective scores.

The DNN architecture used in the experiments was 1799-2048-2048-2048-K, which denoted that the sizes were 1799 (257*7, τ=3) for the input layer, 2048 for three hidden layers, and K for the output layer. K is 257 for DNN-1 and 514 for DNN-2, respectively. The number of epoch for each layer of RBM pre-training was 20 while the learning rate of pre-training was 0.0005. For the fine-tuning, learning 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 [24].

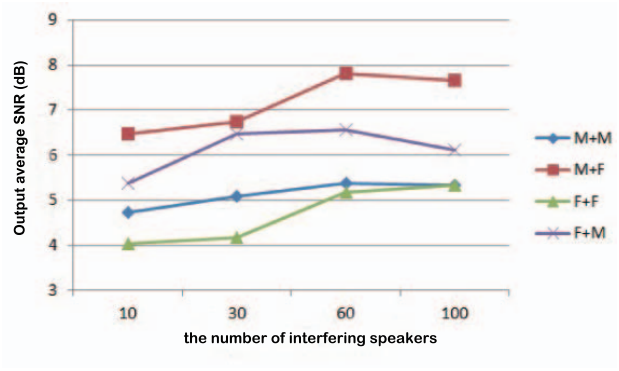

Figure 4: The separation performance (output SNR) comparison of different number of interfering speakers in DNN-2 approach for the target speaker on the test set with four gender combinations.

4.1. Evaluation with different number of interferers

When the experiments were conducted in the semi-supervised mode, the number of interfering speakers in the training stage for predicting the unseen interferer in the separation stage should be determined. Fig. 4 gives a separation performance (in terms output SNR) comparison of different number of interfering speakers for training in the DNN-2 approach for the target speaker on the test set with four gender combinations (target + interferer), namely male and male (M+M), male and female (M+F), female and female (F+F), female and male (F+M). The number of interferers was set to 10, 30, 60, and 100 while the corresponding size of training set was from 10 hours to 100 hours. The output SNR performance was averaged across different input SNRs and different interferers for each gender combination. It was observed that using an adequate size of interferers the trained DNN can well predict an unseen interferer in the separation stage due to the powerful modeling capability of DNN. The best performance was achieved when the number of interferers was 60, which was set as a default for the following experiments. This could be explained as too few interferers could not well predict the unseen interferer while too many speakers might increase the confusion between the target and interferers. Moreover separation of the male target seems to be easier than separating the female target. Meanwhile, separation from the mixture with two different genders could yield much better results than the mixture with the same gender.

4.2. Evaluation of DNN-1 and DNN-2

Fig. 5 shows a STOI comparison of DNN-1 and DNN-2 on the test set for the male (M) or female (F) target speaker under different input SNRs. Noted that the STOI performance was averaged across 25 male and 25 female interferers. The perfor-

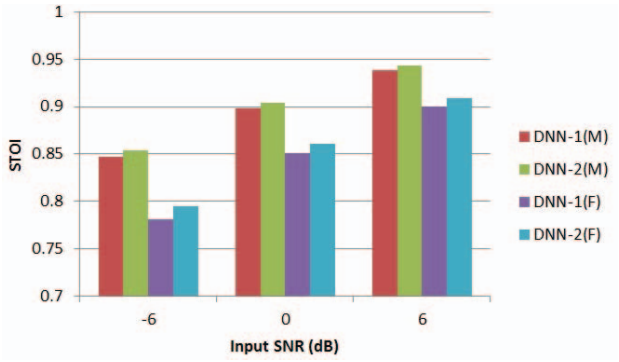

Figure 5: The separation performance (STOI) comparison of DNN-1 and DNN-2 for the male (M) or female (F) target speaker under different input SNRs on the test set.

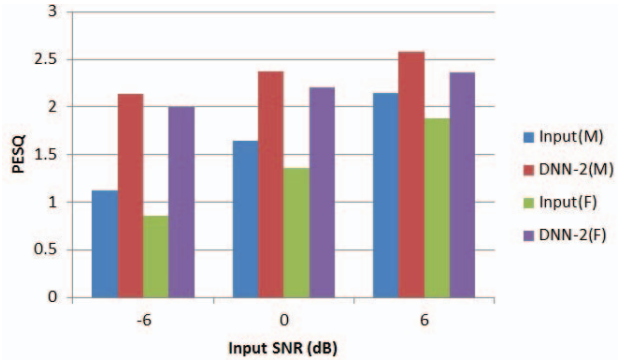

Figure 6: The separation performance (PESQ) comparison of input mixture and DNN-2 approach for the male (M) and female (F) interferers on the test set.

mances of DNN-2 were consistently better than that of DNN-1 for all SNR levels, which confirmed that DNN-2 had better generalization capacity. Similar to Fig. 4, the STOI performance of the male target was always better than that of the female target. And the performance gain of DNN-2 over DNN-1 was more significant for the female target.

Another benefit from DNN-2 is that the interfering speaker can be also separated. Fig. 6 lists a PESQ comparison of the input mixture and the DNN-2 approach for the male (M) and female (F) interferers on the test set. The PESQ performance was averaged across the interferers with the same gender. Obviously, DNN-2 yielded a very significant improvements over the unprocessed input mixture which implied that the unseen interferers could also be well separated from the mixture. The performance gap among different input SNRs of DNN-2 was much smaller than that in the original mixtures, which indicated that DNN-2 approach is more effective under lower SNRs. For example, the PESQ improvement from 1.1 to 2.1 was observer for male interferers at SNR=-6dB while the increment was from 2.1 to 2.55 at SNR=6dB.

Finally, the spectrograms of an utterance example are illustrated in Fig. 7. Fig. 7(a) is the spectrogram of mixed utterance with a female target and a male interferer at 0 dB. Fig. 7(b) is the spectrogram of female target while Fig. 7(c) corresponds to the male interferer. Fig. 7(d) is the spectrogram of DNN-1 separated female target. Fig. 7(e) and (f) are the spectrograms of DNN-2 separated female target and male interference, respectively. For speech separation of the target speaker, both DNN-2 and DNN-1 generated similarly good results which were very close to the reference speech. Another interesting observation was that although there was no information about the interferer, we could still obtain a good separation result of the unseen interferer, which further confirmed that our proposed DNN is effective in predicting unseen interferers by using multiple interfering speakers is training. Furthermore, DNN-2 demonstrated the potential of separating the target speaker from the interferer even the mixed speech is corrupted with noises which is quite common in real applications.

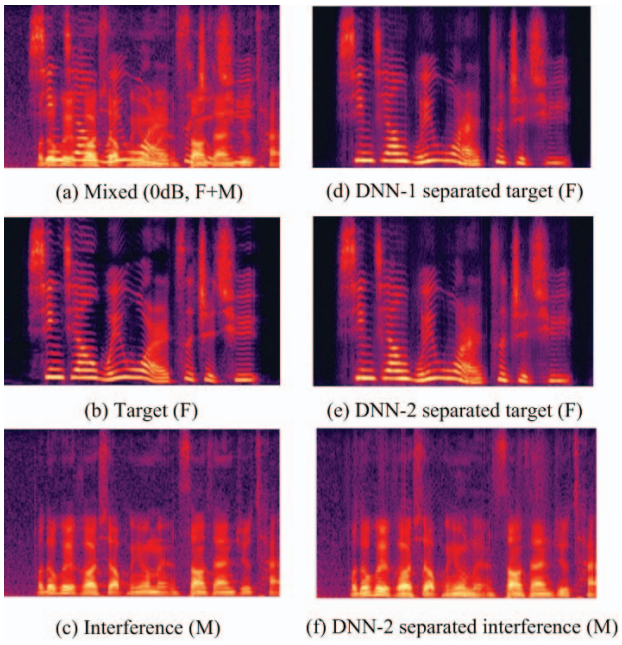

Figure 7: Illustration of spectrograms for (a) input mixture with a female target and a male interferer at 0dB SNR, (b) the female target, (c) the male interference, (d) DNN-1 separated female target, (e) DNN-2 separated female target, (f) DNN-2 separated male interferer.

5. Conclusion and Future Work

In this paper, we have presented a novel architecture of DNN for separating speech of both the target and the interfering speaker. With the additional requirements of predicting the speech feature of the interesting speaker we believe the proposed DNN-2 is more powerful than the baseline DNN-1 in speech separation. In the semi-supervised mode, it demonstrates a better generalization capacity for separating the target speaker while the separated interference can be used for developing other algorithm and applications. Our ongoing research includes extending to single mixtures with more than two speakers, and separating multiple target speakers using one or more DNNs.

6. Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grants No. 61305002 and the Programs for Science and Technology Development of Anhui Province under Grants No. 13Z02008-4 and No. 13Z02008-5.

7. 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, "CLOSEła 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] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

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

[17] 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," *Submitted to ICSP 2014*.

[18] 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.

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

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

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

[22] 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.

[23] 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.

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