Speech Separation of A Target Speaker Based on Deep Neural Networks

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

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

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

Abstract

This paper proposes a novel data-driven approach based on deep neural networks (DNNs) for single-channel speech separation. DNN is adopted to directly model the highly non-linear relationship of speech features between a target speaker and the mixed signals. Both supervised and semi-supervised scenarios are investigated. In the supervised mode, both identities of the target speaker and the interfering speaker are provided. While in the semi-supervised mode, only the target speaker is given. We propose using multiple speakers to be mixed with the target speaker to train the DNN which is shown to well predict an unseen interferer in the separation stage. Experimental results demonstrate that our proposed framework achieves better separation results than a GMM-based approach in the supervised mode. More significantly, in the semi-supervised mode which is believed to be the preferred mode in real-world operations, the DNN-based approach even outperforms the GMM-based approach in the supervised mode.

Index Terms: single-channel speech separation, supervised mode, semi-supervised mode, deep neural networks

1. Introduction

Speech separation aims to separate the voice of each speaker when multiple speakers talk simultaneously, which 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 cochannel) speech separation. Based on the information used in cochannel speech separation, the algorithms can be classified into two categories: unsupervised and supervised approaches. In unsupervised approaches, speaker identities and the reference speech of each speaker for pre-training are not available, while the information of both target and interfering speakers is provided in supervised approaches.

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 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].

In this study, inspired by our recent work on speech enhancement based on deep neural networks (DNNs) [17], we propose to solve the separation problem in Eq. (1) in an alternative way. DNN is adopted to directly model the highly nonlinear relationship of speech features between 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. Our proposed approach avoids the difficult inference based on Eq. (1) using complex models for both target and interfering speakers. As a supervised approach, our experiments show that DNN-based separation achieves significantly better performance than GMM-based separation in [14] due to the powerful modeling capability of DNN. To further verify the effectiveness of our DNN-based approach in a more realistic scenario, namely the semi-supervised mode where only the target speaker information (training data) is given, we propose using multiple speakers to be mixed with the target speaker to train the DNN which is shown to well predict an unseen interferer in the separation stage. More significantly, our DNN-based approach in the semi-supervised mode even outperforms the GMM-based approach in the supervised mode.

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.

978-1-4799-2186-7/14/$31.00 ©2014 IEEE

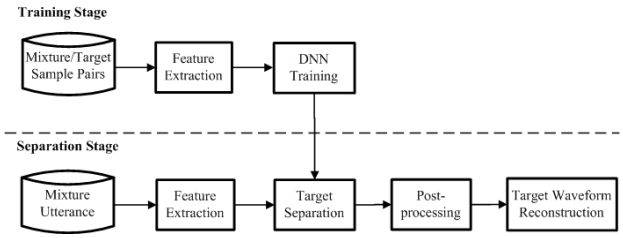

Figure 1: Overall development flow and architecture.

2. System Overview

The overall flowchart of our proposed speech separation system of a target speaker 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 target speaker. 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, followed by a post-processing. 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 [17]. In the next section, the details of DNN training and post-processing are elaborated.

3. DNN-based Speech Separation

In this work, 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, which is shown in Fig. 2. The log-power spectral features can offer 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 levels inherently.

Training of DNN consists of two key steps: unsupervised pre-training and supervised fine-tuning. The pre-training treats each consecutive pair of layers as a restricted Boltzmann machine (RBM) while the parameters of RBM are trained layer by layer with the approximate contrastive divergence algorithm [18]. 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:

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

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. κ is the regularization weighting coefficient to avoid over-fitting. The objective function is optimized using back-propagation procedure with a stochastic gradient descent method in mini-batch mode of N sample frames. Based on our preliminary experiment, we observe that the estimated clean speech has a muffling effect when compared with reference clean speech. To alleviate this problem, global variance equalization (GVE), as a post-processing, is used to further enhance the speech region of the target speaker and suppress the residue of the interferer simultaneously. In GVE, a dimension-independent global equalization factor β can be defined as:

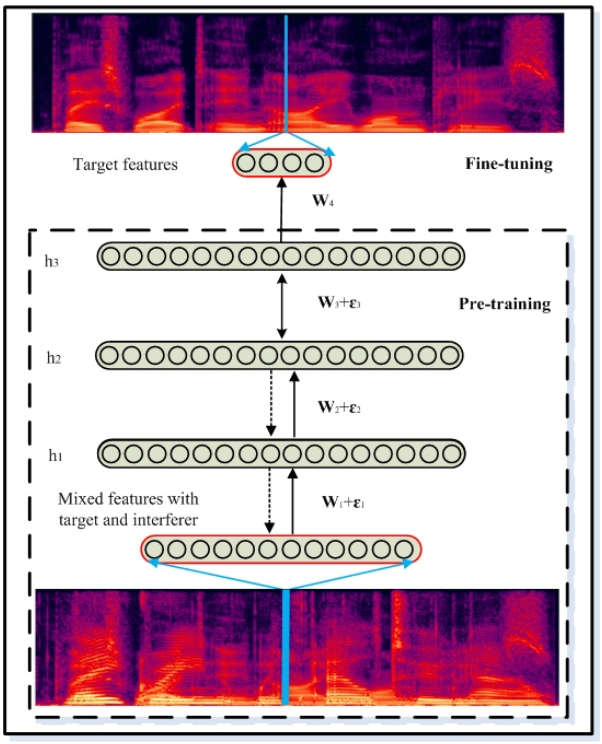

Figure 2: DNN for speech separation.

$$\beta = \sqrt{\frac{GV_{\mathrm{ref}}}{GV_{\mathrm{est}}}} \qquad (3)$$

where GV_{ref} and GV_{est} are the dimension-independent global variance of the reference clean features and the estimated clean features, respectively. Then the post-processing is:

$$\tilde{\boldsymbol{x}}_n^{\mathrm{t}} = \beta\hat{\boldsymbol{x}}_n^{\mathrm{t}} \qquad (4)$$

where $\tilde{\boldsymbol{x}}_n^{\mathrm{t}}$ is the final estimated clean speech feature vector.

To investigate the effectiveness of the proposed DNN-based separation approach, experiments in both supervised and semi-supervised modes are designed. One case is a mixture consists of one target and only one interferer, denoted as 1+1 mode. Then each mixture utterance for training of DNN is synthesized by adding the randomly selected segment of the interferer with a specified SNR to the utterance of the target speaker. In the separation stage, only the mixture with the same target and interferer is tested in a supervised manner. The other case is a

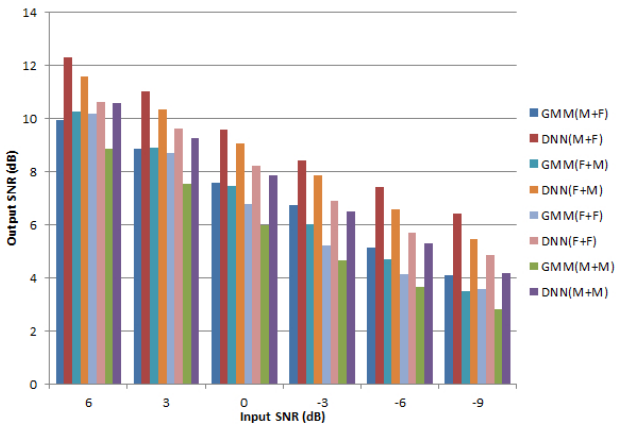

Figure 3: Output SNR comparison of different approaches with four gender combinations in the 1+1 supervised mode.

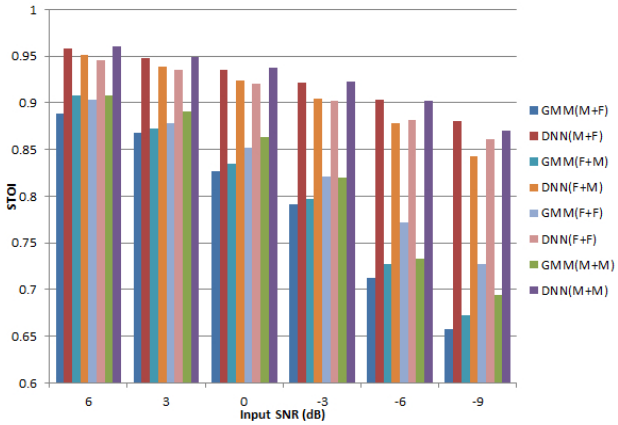

Figure 4: STOI comparison of different approaches with four gender combinations in the 1+1 supervised mode.

mixture consists of one target and N possible interferers, denoted as 1+N mode. Then each mixture utterance for training of DNN is synthesized by adding the randomly selected segment of one interferer from N possible interferers with a specified SNR to the utterance of the target speaker. In the separation stage, if the interferer in the mixture is still among the N possible interferers used in the training stage, then the separation is in a supervised manner. Otherwise, the separation is in a semi-supervised manner with an unseen interferer.

4. Experiments and Result Analysis

Our experiments were conducted on the Speech Separation Challenge (SSC) corpus [19]. 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 SNRs ranging from -10 dB to 10 dB with an increment of 1 dB. We use the test set of the SSC corpus with two-speaker mixtures at SNRs from -9 dB to 6 dB with an increment of 3 dB for evaluation. Note that the mixture utterances were the same across different SNRs. Obviously, the mixtures in the training set have a good SNR coverage for the test set. The method in [14] is adopted for performance comparison with our DNN approach, which is denoted as "GMM" approach in the following experiments.

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 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 two measures, namely output SNR [14] and short-time objective intelligibility (STOI) [20] believed to be highly correlated to speech intelligibility.

The DNN architecture used in the experiments was 1799-2048-2048-2048-257, which denoted that the sizes were 1799 (257*7, τ=3) for the input layer, 2048 for three hidden layers, and 257 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 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

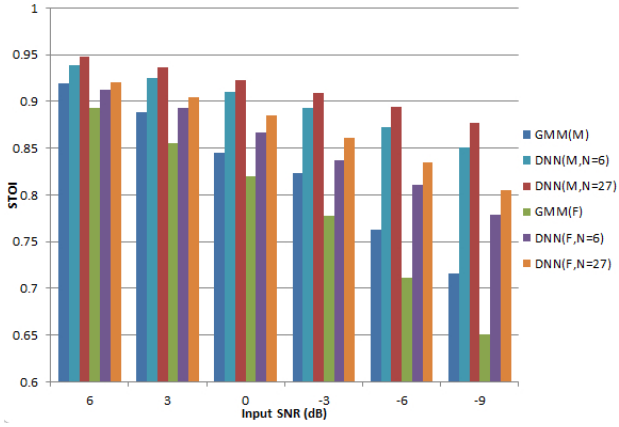

Figure 5: STOI comparison of different approaches with the female (F) and the male (M) targets in the 1+N supervised mode.

normalized to zero mean and unit variance. Other parameter settings can refer to [21].

4.1. Evaluation in 1+1 mode

In the 1+1 supervised mode, information of both the target and interferer is provided in advance and there is only one interferer. As the training of each DNN with one target and one interferer was time-consuming, 16 combinations of targets and interferers were randomly selected for training and evaluation, which were equally assigned for four possible gender combinations, namely female and female (F+F), male and male (M+M), female and male (F+M), male and female (M+F). For each combination, about 30 hours of mixed speech were synthesized by the target and interferer for corresponding DNN training. Fig 3 gives a performance (output SNR) comparison of different separation approaches with four gender combinations in 1+1 supervised mode. First, all DNN systems significantly improved the output SNR over the GMM systems across different input SNRs (more than 2 dB improvement in the best case). Second, the output SNRs for different gender combinations in both GMM and DNN approaches roughly followed a certain trend across different input SNRs, namely, monotonically decreased in the

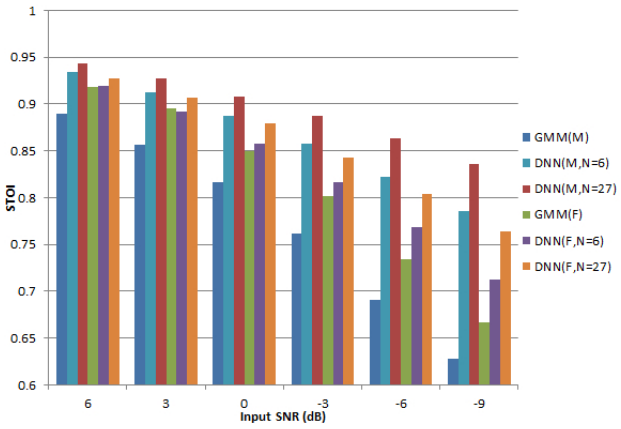

Figure 6: STOI comparison of different approaches with the female target (F) and the male target (M) in the 1+N semi-supervised mode.

order of M+F, F+M, F+F, and M+M. Fig 4 presents the corresponding STOI comparison. Not surprisingly, DNN still consistently outperformed GMM. However, it was interesting that our proposed DNN approach was more robust to the decreased input SNRs than GMM. Even in the -9 dB condition, STOI of DNN could achieve about 0.85, which was comparable to that of the GMM approach at 3 dB.

4.2. Evaluation in 1+N mode

In the 1+N mode, 1 target and N interferers were used to generate the mixed speech with two speakers in the training stage. In the test stage of separating the target, if the interferer is one of the N interferers in the training stage, then it is still a supervised mode. Otherwise, it is a semi-supervised mode with an unknown interferer. To test the effect of N, experiments on N=6 and N=27 were conducted. The data amount of mixed speech synthesized as the training set for N=6 and N=27 were about 30 hours and 140 hours, respectively. Training of DNN with such an amount of data was time-consuming. So only one female target and one male target were selected, and all the mixtures with those two targets on the test set were used for evaluation in the following experiments. Fig. 5 shows an STOI comparison of different approaches with the female target (F) and the male target (M) in the 1+N supervised mode. We can observe that increasing N with more training data could always improve STOI in the proposed DNN approach. Similar to Fig 4, the STOIs of DNN were much better than those of GMM even with more confusing interferers included. Fig. 6 lists an STOI comparison of the different approaches with the female target (F) and the male target (M) in the 1+N semi-supervised mode. Note that the results for GMM in Fig. 6 are still in a supervised mode. Similar observations can also be made as those in Fig. 5. There was only one exception that DNN(F,N=6) at 3 dB generateed worse STOI than GMM. Overall, the DNN approach with N=27 achieved consistently the best separation performance. These results were very encouraging as our DNN approach without any information about the interferers could beat the conventional GMM approach with information of both the target and the interferer. This confirms that using many interferers in training DNN can well predict an unseen interferer in the separation stage due to the powerful modeling capability of DNN.

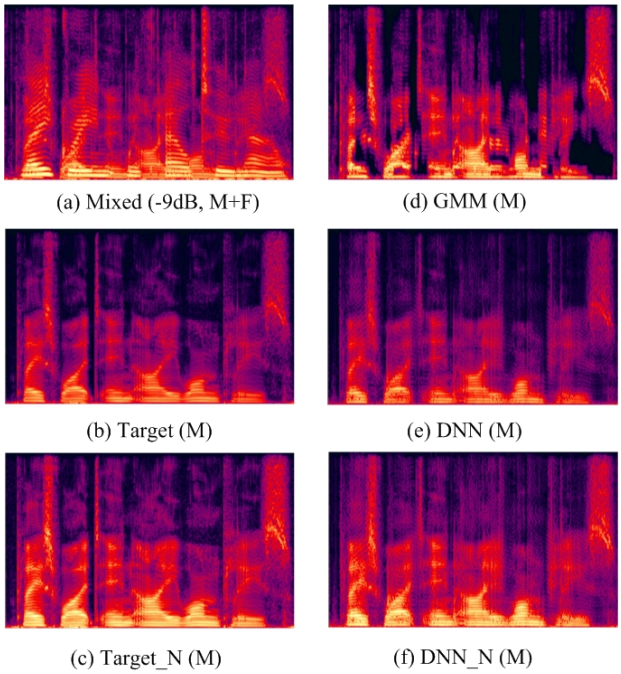

Figure 7: Illustration of spectrograms for separating the target male utterance from the mixed utterance with a female interferer in the semi-supervised 1+N mode (N=27).

Finally, the spectrograms of an utterance example are illustrated in Fig. 7 with Fig. 7(a) for a mixed utterance with a male target and a female interferer at -9 dB and Fig. 7(b) for the target male. While Fig. 7(c) is for a corresponding version with energy normalization as in [14], which is used as a reference for Fig. 7(d) using a GMM approach where energy normalization should be performed. Fig. 7(e) is the spectrogram of our proposed approach in the semi-supervised 1+N mode (N=27). To give a fair comparison with Fig. 7(d), the normalized version of our result is also shown in Fig. 7(f). Obviously, our results are closer to the target reference than that of the GMM approach. It is also interesting to note that no interferer information is given.

5. Conclusion and Future Work

In this paper, we present a novel DNN-based approach to single-channel speech separation. We demonstrate its effectiveness over state-of-the-art approaches of separating a single target speaker from mixtures of two voices in both the supervised and semi-supervised modes. With more training speech data from interfering speakers, the performance in the semi-supervised mode can even surpass that of the GMM approach in the supervised mode. Ongoing future work includes extending the separation of a single target from the mixture utterance with more than one interfering speaker and investigating the separation of multiple target speakers using a single DNN.

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, China 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. pp.1-11, 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] 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.

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

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

[20] 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.

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