Deep Neural Network Based Speech Separation for Robust Speech Recognition

Tu Yanhui[1], *Du Jun*[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

{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 without using any prior information about the interfering speaker. DNN is adopted here to directly model the highly nonlinear relationship between speech features of the mixed signals and the two competing speakers. Experimental results on a monaural speech separation and recognition challenge task show that the proposed DNN framework 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. Furthermore, as a preprocessing step in the testing stage for robust speech recognition, our speech separation approach can achieve significant improvements of the recognition accuracy over the baseline system with no source separation.

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, such as speech communication and automatic speech recognition. In this study, we focus on separating two competing 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 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, 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 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 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] 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 [10] investigates the phone-level dynamics using HMMs [9]. For FHMM-based speech separation, 2-D Viterbi algorithms and approximations have been used to perform 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].

One recent work [18] uses deep neural networks (DNNs) to solve the separation problem in Eq. (1) in an alternative way. DNN was adopted to directly model the highly nonlinear relationship between speech features of a target speaker and the mixed signals. Eq. (1) plays the role of simulating a large amount of the mixed speech and target speech pairs for DNN training, given the speech sources of the target speaker and interfering speaker. 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. This proposed framework avoids specifying 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 [15] due to the powerful modeling capability of DNN. With this newly defined objective function aiming at minimizing the mean squared error between the DNN output and the reference clean features of both speakers, the proposed dual-output objective function leads to an improved generalization capacity to unseen interferers for separating the target speech signals. Meanwhile, without any prior information from the interferer, the interference speech can also be well separated for developing new algorithms and applications. For evaluating our proposed approach, both speech separation and recognition experiments were conducted on a monaural speech separation and recognition challenge task initiated in 2006 [19, 20, 21], and very promising separation results and improved recognition accuracies were achieved with the proposed DNN approach.

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

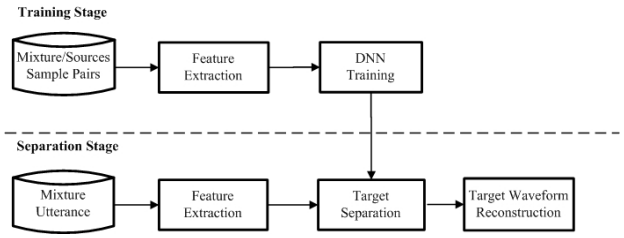

Figure 1: Development flow for speech separation system.

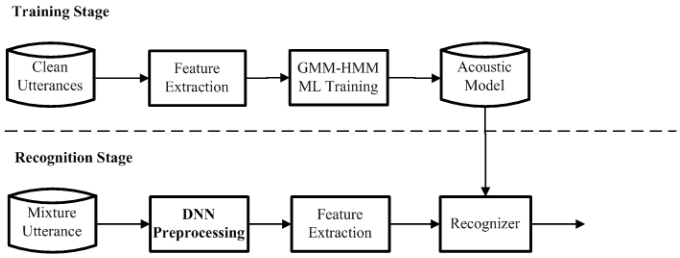

Figure 2: Development flow for speech recognition system.

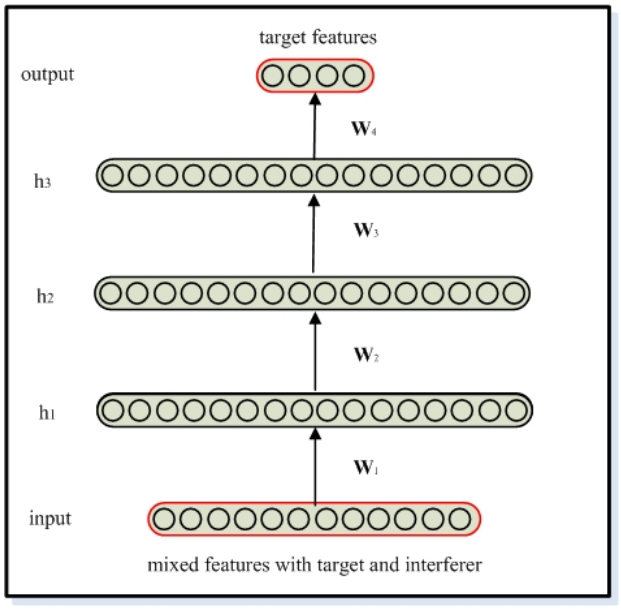

Figure 3: DNN-1 architecture.

The remainder of the paper is organized as follows. In Section 2, we give a system overview. In Section 3, we introduce 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 speech separation system and recognition system 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 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 [22]. 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 using Mel-frequency cepstral coefficients (MFCCs) under the maximum likelihood (ML) criterion. In the recognition stage, the mixture utterance is first preprocessed by the DNN model to extract the speech waveform of the target speaker. Then the normal feature extraction and recognition is conducted. In the next section, the detail of two types of DNN architectures are elaborated.

3. DNN-based Speech Separation

3.1. DNN-1 for predicting the target

In [18], DNN was 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. 3. 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 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 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 a full coverage of a large number of the SNR levels in the training stage.

Training of DNN consists of unsupervised pre-training and supervised fine-tuning. Pre-training treats each consecutive pair of layers as a restricted Boltzmann machine (RBM) [23] while the parameters of RBM are trained layer by layer with the approximate contrastive divergence algorithm [24]. For supervised fine-tuning, we aim at minimizing the 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**.

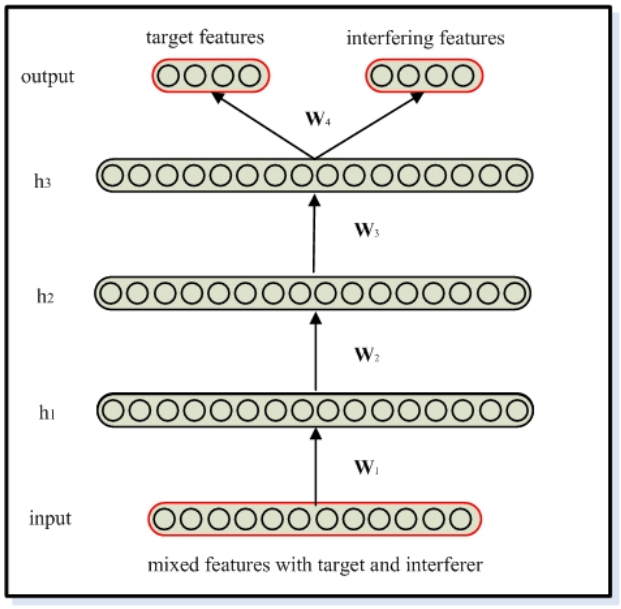

Figure 4: DNN-2 architecture.

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

Next we designed a new DNN architecture for speech separation which is illustrated in Fig. 4. The main difference from Fig. 3 is that the new DNN can predict both the target and interference in the output layer which is denoted as **DNN-2**. Pre-training of DNN-2 was exactly the same as that of DNN-1 while the supervised fine-tuning was conducted by jointly minimizing the mean squared error between the DNN output and the reference clean features of both the target and interference:

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

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 [15], 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 [18], 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 mainly focus 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 unseen interferers in the separation stage [18].

4. Experiments

Experiments were conducted on the SSC (Speech Separation Challenge) corpus [19] for recognizing a few keywords from simple *target* sentences when presented with a simultaneous *masker* sentence with a very similar structure [20]. All the training and test materials were drawn from the GRID corpus [25]. There are 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) from -9dB to 6dB with an increment of 3dB. For training 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.

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 transfore (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 a short-time objective intelligibility (STOI) [26], and perceptual evaluation of speech quality (PESQ) [27]. STOI is shown to be highly correlated to human speech intelligibility while PESQ has a high correlation with subjective listening scores.

The DNN architecture used in all 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 fine-tuning, the 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 [28].

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 derivatives. Each word was modeled by a whole-word left-to-right HMMs with 32 Gaussian mixtures per state. The number of states for each word can be referred to [20].

4.1. Experiments on speech separation

Fig. 5 shows a STOI comparison of DNN-1 and DNN-2 for one male (M) or female (F) target speaker under different input SNRs in the semi-supervised mode. The number of interfering speakers in the training stage was set to 27. The data amount of mixed speech synthesized as the training set was about 140 hours for each DNN of the target speaker. All the mixtures with those two targets on the test set were used for evaluation. The performances of DNN-2 were consistently better than those of DNN-1 at all SNR levels, which confirms that DNN-2 has better generalization capacity than DNN-1.

Fig. 6 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 50 hours of mixed speech for each target speaker. Totally 34 DNNs were trained for all target speakers.

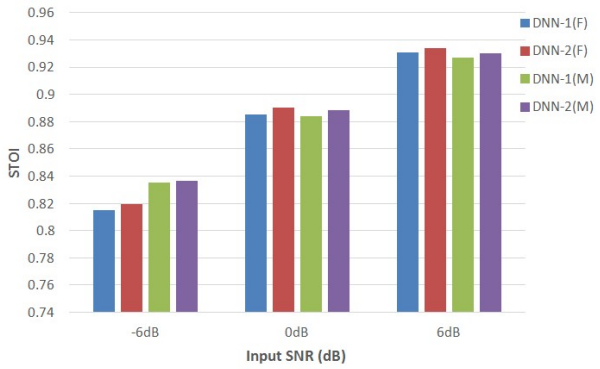

Figure 5: Separation performance (STOI) comparison of DNN-1 and DNN-2 for one male (M) and one female (F) target speaker under different SNRs in semi-supervised mode.

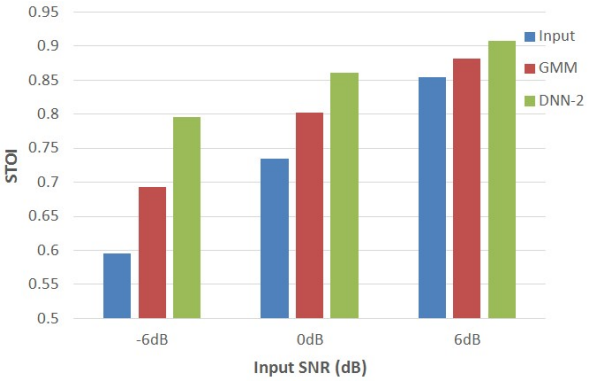

Figure 6: Separation performance (STOI) comparison of different approaches averaged across all 34 testing target speakers.

Noted that both seen and unseen interfering speakers (compared with those in the training set) were included on the test set for evaluation. The method in [15], denoted as "GMM" approach, was adopted for a performance comparison with our DNN approach. Obviously, DNN-2 yielded a very significant improvements of STOI performance over both the unprocessed input mixture and GMM approach. The performance gaps among different input SNRs for DNN-2 was much smaller than those in the GMM approach, which indicates that the DNN-2 approach is more effective under lower SNRs. For example, the STOI improvement from 0.69 to 0.8 was observed from GMM to DNN-2 at SNR=-6dB while the increment was only from 0.88 to 0.91 at SNR=6dB. The corresponding PESQ performance comparison given in Fig. 7 can also draw similar observations.

4.2. Experiments on robust speech recognition

Finally, the effectiveness of the proposed DNN-based separation approach is further verified for speech recognition. The same configurations for DNN training as in Fig. 6 were adopted. In Table 1, we report the performance (word accuracy in %) comparison of the baseline system and the DNN-2 preprocessed system averaged across all female and male target speakers on the test set. Very promising results were achieved using DNN-2 preprocessing under different SNRs for both female and male target speakers, e.g., the relative word error rate reduction was up to 83% at SNR=6dB and at least 54% at SNR=-9dB.

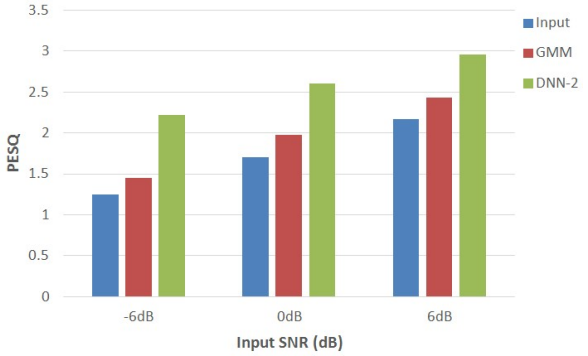

Figure 7: Separation performance (PESQ) comparison of different approaches averaged across all 34 testing target speakers.

Table 1: Performance (word accuracy in %) comparison of baseline system and DNN-2 preprocessed system averaged across all female and male target speakers on the test set.

	Baseline		DNN-2	
Input SNR (dB)	Female	Male	Female	Male
6	45.69	52.20	90.48	92.11
3	31.03	36.83	83.93	86.56
0	21.26	24.39	78.57	82.11
-3	11.78	15.61	71.43	73.68
-6	9.20	11.22	65.48	65.43
-9	7.47	8.54	57.14	59.04

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 sppech 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 proposed approach also shows the effectiveness for robust speech recognition as a preprocessing step. Our ongoing research includes extending to single mixtures with more than two speakers, separating multiple target speakers using one or more DNNs, and further improving the recognition accuracy with other techniques.

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] Lawrence R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, 77 (2): 257C286. doi:10.1109/5.18626. February 1989

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

[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] 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 Proc. ICSP*, 2014.

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

[20] 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.

[21] 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.

[22] 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.

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

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

[25] 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.

[26] 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.

[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.

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