A Regression Approach to Single-Channel Speech Separation Via High-Resolution Deep Neural Networks

Jun Du, Yanhui Tu, Li-Rong Dai, and Chin-Hui Lee, *Fellow, IEEE*

***Abstract*—We propose a novel data-driven approach to single-channel speech separation based on deep neural networks (DNNs) to directly model the highly nonlinear relationship between speech features of a mixed signal containing a target speaker and other interfering speakers. We focus our discussion on a semisupervised mode to separate speech of the target speaker from an unknown interfering speaker, which is more flexible than the conventional supervised mode with known information of both the target and interfering speakers. Two key issues are investigated. First, we propose a DNN architecture with dual outputs of the features of both the target and interfering speakers, which is shown to achieve a better generalization capability than that with output features of only the target speaker. Second, we propose using a set of multiple DNNs, each intending to be signal-noise-dependent (SND), to cope with the difficulty that one single general DNN could not well accommodate all the speaker mixing variabilities at different signal-to-noise ratio (SNR) levels. Experimental results on the speech separation challenge (SSC) data demonstrate that our proposed framework achieves better separation results than other conventional approaches in a supervised or semisupervised mode. SND-DNNs could also yield significant performance improvements over a general DNN for speech separation in low SNR cases. Furthermore, for automatic speech recognition (ASR) following speech separation, this purely front-end processing with a single set of speaker-independent ASR acoustic models, achieves a relative word error rate (WER) reduction of 11.6% over a state-of-the-art separation and recognition system where a complicated joint back-end decoding framework with multiple sets of speaker-dependent ASR acoustic models needs to be implemented. When speaker-adaptive ASR acoustic models for the target speakers are adopted for the enhanced signals, another 12.1% WER reduction over our best speaker-independent ASR system is achieved.**

***Index Terms*—Deep neural network, divide and conquer, dual outputs, robust speech recognition, speech separation.**

I. Introduction

SOURCE separation [1]–[3] is a long-standing classical signal processing problem about separating individual signals from multiple sources received in a mixed mode. Speech separation [4]–[6], to be specific, aims at singling out the voice of each speaker in the mixed speech with multiple speakers talking at about the same time. Single-channel speech separation [7]–[9] often refers to the situation that the mixed speech signals are recorded with a single microphone. Assumptions about each of the signal components already mixed are often required in order to obtain a satisfactory separation performance [10], [11]. One broad class of single-channel speech separation is the so-called computational auditory scene analysis (CASA) [12], usually performed in an unsupervised mode referring to the situation that speaker identities and the reference speech for each speaker are not available in the training stage. CASA-based approaches [13]–[17] use a set of 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 [16], pitch and amplitude modulation were adopted to separate the voiced portions of co-channel speech [18], [19]. In [17], unsupervised clustering was used to categorize speech regions into two speaker groups by maximizing the ratio of the between-cluster and within-cluster distances. Recently, a data-driven approach [20] attempts to separate the underlying clean speech segments by matching each mixed speech segment against a composite training segment.

On the other hand in a supervised mode, in which some information of both the target and the interfering speakers is provided, 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 [10] adopted a factorial hidden Markov model (FHMM) to describe a speaker, and the estimated sources were used to generate a binary mask. To further impose temporal constraints on speech signals for separation, the work in [21] investigated the phone-level dynamics using HMMs [22]. For FHMM-based speech separation, 2-D Viterbi algorithms and approximations have been used to perform the inference [23]. In [9], FHMM was adopted to model vocal tract characteristics for detecting pitch to reconstruct speech sources. In [8], [11], [24], Gaussian mixture models (GMMs) [25], [26] were employed to model speakers, and minimum mean squared error (MMSE) or maximum a *posteriori* (MAP) based estimators were used to recover the speech signals. The factorial-max vector quantization model was also

Manuscript received February 15, 2016; accepted April 19, 2016. Date of publication April 27, 2016; date of current version May 27, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yunxin Zhao.

J. Du, Y. Tu, and L.-R. Dai are with the National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei 230027, China (e-mail: jundu@ustc.edu.cn; tuyanhui@mail.ustc.edu.cn; lrdai@ustc.edu.cn).

C.-H. Lee is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: chl@ece.gatech.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2558822

used to infer the mask signals in [27]. Other popular approaches include nonnegative matrix factorization (NMF) based models [7].

Recently, deep learning techniques become increasingly popular in many speech research areas, e.g., speech recognition [28]–[30], speech enhancement [31], [32], speech synthesis [33], [34]. In this study, inspired by our recent work on speech enhancement based on deep neural networks (DNNs) [32], [35], we propose to solve the separation problem in Eq. (1) by adopting DNNs to directly model the highly non-linear relationship of speech features from the mixed signals to a target speaker [36] and possibly an interference speaker as well. Eq. (1) plays the role of synthesizing a large amount of mixed speech for DNN training, given the speech sources of the target speaker and interfering speakers. Our proposed approach avoids the difficulty of using complex but inaccurate model assumptions for both the target and interfering speakers based on Eq. (1).

As a supervised approach, our experiments show that DNN-based separation achieves a superior performance to GMM-based separation in [24] due to the powerful modeling capability of the nonlinear mapping functions implied by DNNs. More recently, several representative approaches via deep learning [31], [37]–[44] are also proposed to solve the single-channel speech separation. One category discusses the segregation of the background noises from the speech, including the noise perturbation for DNN-based speech separation [37], the use of long short-term memory based recurrent neural networks (RNNs) [38] and deep NMF models [39]. Another category, which is more related with our work, aims to segregate the speech from mixed speakers. In [40], [41], DNN and RNN are adopted to estimate each source of the mixed speech with a post-processing using masking technique. The main differences from our approaches are: (i) although our DNN architecture with dual outputs [45] is similar to that in [40], [41], the learning targets are log-power spectra rather than the power spectra used in [40], [41], which is inspired by that the MMSE criterion in the log-domain is more consistent with the human auditory system [46]; (ii) we next 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 in a more realistic scenario, namely the semi-supervised mode where only the target speaker information (characterized by training data) is given. More significantly, our DNN-based approach in the semi-supervised mode even outperforms the GMM-based approach in the supervised mode. Another related work is the generative stochastic network (GSN) based speech separation [42]–[44] via a hybrid generative-discriminative training objective. The main difference from our approach is the use of the generative term, which requires a development set to tune the weighting factor. Compared with all these deep learning approaches, our approach is also innovative in adopting a divide and conquer strategy to design signal-noise-dependent DNNs (SND-DNNs) with a detailed resolution [47] by considering that a single general DNN might not be able to well accommodate all the variabilities at a wide range of signal-to-noise-ratio (SNR) levels. Two SND-DNNs are trained to cover the mixed speech with positive 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 used in the second-pass SND-DNN based separation.

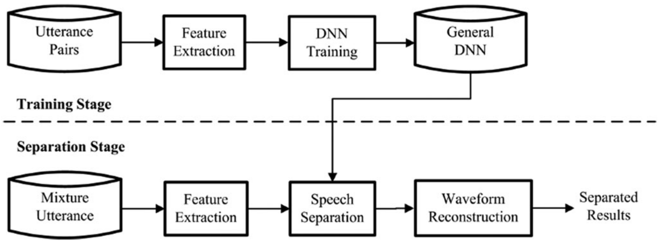


Fig. 1. Development flow for DNN-based separation system.

Finally, a comprehensive series of experiments are conducted for speech separation, especially with a larger scale of training data (typically about 100 hours) in comparison to other approaches [40]–[44] which is crucial for improving the separation performance. Its effectiveness has also been verified for robust speech recognition [48]. The evaluation results on the speech separation challenge (SSC) corpus [49] show that the proposed SND-DNNs approach significantly outperforms the general DNN approach [48] in terms of both separation and recognition performance. Furthermore, our purely front-end only pre-processing method achieves significant performance improvements over the best system in the competition [50], [51] and a comparable performance with the recent work in [52], where a complicated joint decoding framework or/and DNN-based acoustic modeling are implemented in the back-end.

The rest of the paper is organized as follows. In Section II, we first give an overview of our proposed speech separation and recognition systems. In Section III, DNN-based speech separation is described in detail. In Section IV, SND-DNNs based approach is elaborated. In Section V, we report experimental results on speech separation and speech recognition. Finally we summarize our findings in Section VI.

II. System Overview

An overall flowchart of the general DNN based speech separation system is illustrated in Fig. 1. In the training stage, the general DNN as a regression model is trained by using log-power spectral features from pairs of the mixed signal and the individual sources. Note that in this work we only consider the case of two speakers in the mixed signals, namely one target speaker and one 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 [32]. As an improved version, the SND-DNNs based system is given in Fig. 2. The main difference from the general DNN based system is that two SND-DNNs, namely positive and negative DNNs, are trained using mixture

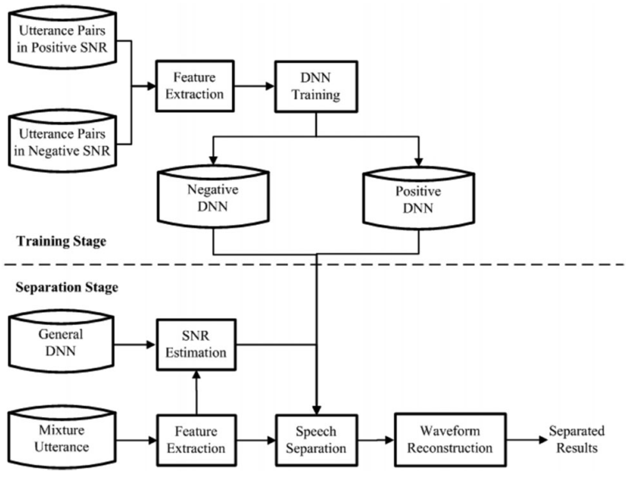


Fig. 2. Development flow for SND-DNNs based separation system.

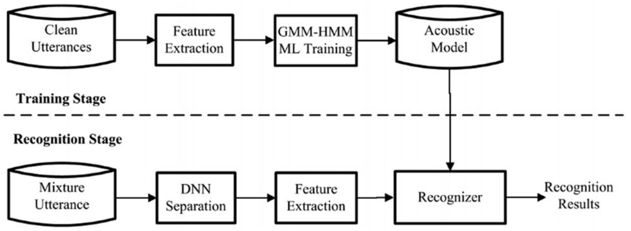


Fig. 3. Development flow for speech recognition system.

utterances with positive and negative SNRs, respectively. In the separating stage, we use a general DNN to perform the 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. 3, the development flow of an automatic speech recognition (ASR) system as one application of our separation approaches is introduced. In the training stage, the acoustic models using Gaussian mixture continuous density HMMs (denoted as GMM-HMMs) or using DNN based HMMs (denoted as DNN-HMMs) [28], [30] are trained from the clean speech of the target speaker using mel-frequency cepstral coefficients (MFCCs) or other features under the maximum likelihood criterion for GMM-HMMs or minimum cross-entropy for DNN-HMMs. In the recognition stage, the mixture utterance is first preprocessed by speech separation based on DNN or SND-DNNs to extract the speech waveforms of the target speaker. Then the conventional feature extraction and recognition follow. In this study, the recognition experiments are only conducted for GMM-HMMs with MFCC features as we mainly focus on the speech separation part as a front-end processing. In the next two sections, the details of DNN and SND-DNNs based approaches are elaborated.

III. DNN Based Speech Separation

A. DNN Architectures: Single Versus Dual Outputs

In this work, DNN is adopted as a regression model to predict the log-power spectral features of the target speaker along with those of interfering speakers given the input log-power

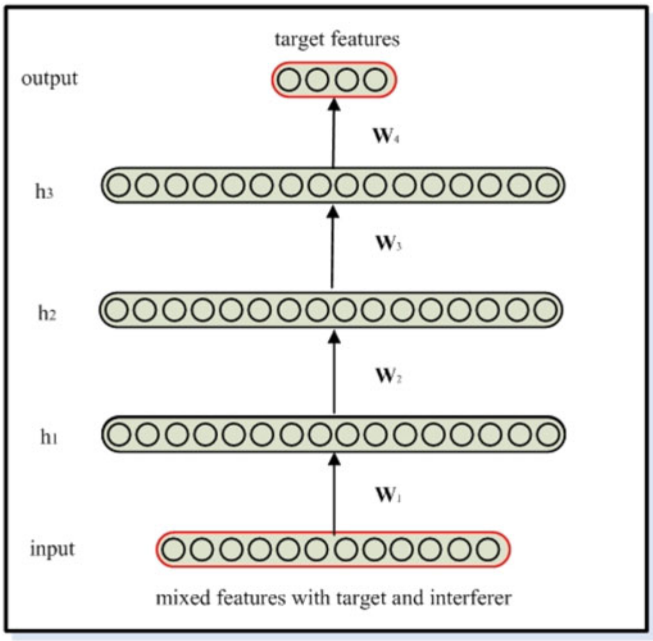


Fig. 4. DNN-1 architecture.

spectral features of mixed speech. Two types of DNN architectures are investigated. One network configuration is a DNN with a single set of output features for the target speaker, denoted as *DNN-1*, which is shown in Fig. 4. We use the log-power spectral features which can offer perceptually relevant speech parameters. The acoustic context information along both the time axis (with multiple neighboring frames) and the frequency axis (with full frequency bins) can be fully utilized by DNN to improve the continuity of the estimated target speech signals while the conventional GMM-based approach does not fully explore the temporal dynamics of speech. As training of this regression DNN requires a large amount of time-synchronized stereo-data of target and mixed speech pairs, the mixed training speech utterances are synthesized by corrupting the speech utterances of the target speaker with interferers at different SNRs (here we consider interfering speech as noise) based on Eq. (1). Note that generalization to different SNR levels in the separation stage can inherently be well addressed by a full coverage of SNR levels in the mixed speech training set.

Training of DNN-1 consists of unsupervised pre-training and supervised fine-tuning. Pre-training treats each consecutive pair of layers as a restricted Boltzmann machine (RBM) [53] while the parameters of RBM are trained layer by layer with the approximate contrastive divergence algorithm [54]. For supervised fine-tuning, we aim at minimizing the mean squared error, E_1, between the predicted DNN output and the reference clean features of the target speaker shown below:

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

where $\hat{\boldsymbol{x}}_n^{\mathrm{t}}$ and $\boldsymbol{x}_n^{\mathrm{t}}$ are the current $n\mathrm{th}$ D-dimensional vectors of estimated and reference clean features of the target speaker,

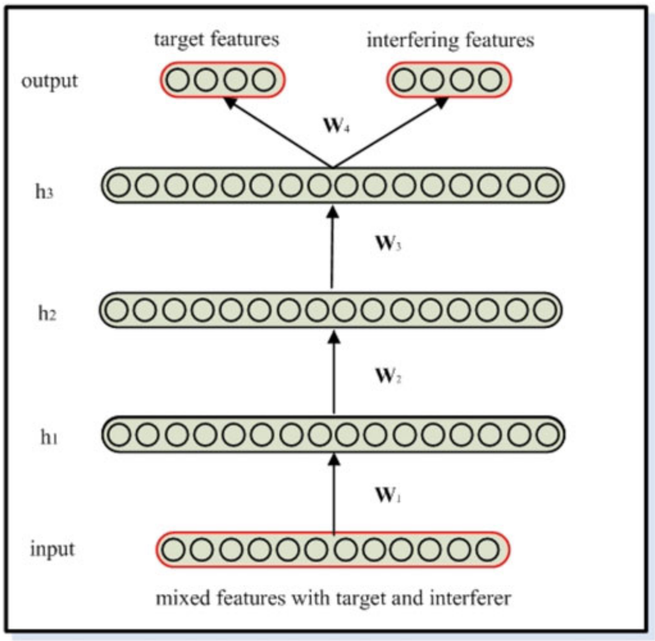


Fig. 5. DNN-2 architecture.

respectively. $\boldsymbol{x}^{\mathrm{m}}_{n\pm\tau}$ is a $D(2\tau+1)$-dimensional vector of the input mixed features with the preceding left and following right τ frames as the acoustic context. $\boldsymbol{W}$ and $\boldsymbol{b}$ denote all the weight and bias parameters of the DNN. The objective function is optimized using back-propagation with a stochastic gradient descent (SGD) method in a mini-batch mode of T sample frames.

The other DNN configuration is with dual outputs, denoted as *DNN-2*, and illustrated in Fig. 5. The main difference from DNN-1 is that DNN-2 predicts both the target and interference at the output layer. Pre-training of DNN-2 is exactly the same as that of DNN-1 while supervised fine-tuning is performed by jointly minimizing the combined mean squared error, E_2, between the estimated DNN output and the reference clean features of both the target and interference speakers as follows:

$$E_2 = \frac{1}{T}\sum_{n=1}^{T}\left(\|\hat{\boldsymbol{x}}^{\mathrm{t}}_n(\boldsymbol{x}^{\mathrm{m}}_{n\pm\tau},\boldsymbol{W},\boldsymbol{b}) - \boldsymbol{x}^{\mathrm{t}}_n\|_2^2\right.$$
$$\left. + \|\hat{\boldsymbol{x}}^{\mathrm{i}}_n(\boldsymbol{x}^{\mathrm{m}}_{n\pm\tau},\boldsymbol{W},\boldsymbol{b}) - \boldsymbol{x}^{\mathrm{i}}_n\|_2^2\right) \tag{3}$$

where $\hat{\boldsymbol{x}}^{\mathrm{i}}_n$ and $\boldsymbol{x}^{\mathrm{i}}_n$ are the nth D-dimensional vectors of the 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 can potentially lead to better generalization than DNN-1 for separating the target speaker. Another benefit from DNN-2 is that the interference speech signal can also be separated as a by-product for developing new advanced algorithms in other applications.

B. Supervised and Semi-Supervised Separation Setups

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 mixture consisting 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 the set of 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. Our definitions of supervised and semi-supervised modes are also corresponding to the speaker dependent (SD) and speaker independent (SI) terms in [44] by considering the information awareness of the interferers. But to avoid the confusion with the same terms for ASR experiments in Table VIII, the supervised or semi-supervised terms will be used by default.

IV. Signal-Noise-Dependent DNNs

So far in our proposed DNN-based speech separation, one single DNN is expected to accommodate all the speaker mixing conditions at different SNRs. A further complication is for the semi-supervised mode in which the interferer is also unseen a general DNN is usually limited in separation capabilities. To circumvent this difficulty, we adopt a divide and conquer strategy to design multiple DNNs with a detailed resolution, each is capable of handling some specific conditions. As a demonstration, we propose the use of signal-noise-dependent DNNs to alleviate the problem of the mixing variabilities caused by different noise levels. As shown in Fig. 2, two SND-DNNs, namely positive DNN and negative DNN, are generated using mixture utterances with positive and negative SNRs, respectively. In the separating stage, the separated target and interference utterances by the general DNN with dual outputs can be used for SNR estimation of the current utterance according to the following equation:

$$\mathrm{SNR} = 10\log\left(\frac{\sum_m x_{\mathrm{t}}^2[m]}{\sum_m x_{\mathrm{i}}^2[m]}\right) \tag{4}$$

where $x_{\mathrm{t}}[m]$ and $x_{\mathrm{i}}[m]$ are the mth samples of the reconstructed target and interference signals in the time domain, respectively. With this estimated SNR level, the corresponding SND-DNNs (positive or negative SNR in this case) can be selected for 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 we could achieve both high model resolution and accurate model selection to be illustrated next. A similar idea is used in [55] to train SNR dependent multilayer perceptrons (MLPs) and use SNR estimation for MLP selection in the context of robust speaker identification.

V. Experiments and Result Analysis

Separation experiments were conducted on the SSC corpus [49] originally designed for recognizing a few keywords embedded in simple *target* utterances but mixed with another simultaneous *masker* utterance by a competing speaker with a very similar structure [50]. All the training and test materials were drawn from the GRID corpus [56]. There were 34 speakers for both training and testing, 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 consisted of two-speaker mixtures at a range of SNRs from -9 dB to 6 dB with an increment of 3 dB. 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 signals at SNR levels 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 was trained using the mixture speech with SNRs ranging from 0 dB to 10 dB. To accommodate for possible errors in SNR estimation, the 0 dB section of the training set was included in both subsets of training speech. Obviously, the mixtures in the training set have a good SNR coverage for the test set. The method in [24], denoted as “GMM” approach in the following experiments, was adopted for separation performance comparisons with the proposed DNN framework.

As for the signal analysis, all waveforms were down-sampled from 25 kHz to 16 kHz, 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 spectral features were used to train DNNs. The separation performance was evaluated using several measures, including output SNR [24], a short-time objective intelligibility (STOI) [57] believed to be highly correlated to speech intelligibility, perceptual evaluation of speech quality (PESQ) [58] with a high correlation to subjective scores, and the corresponding recognition accuracy. The DNN-1 architecture used in all experiments was 1799-2048-2048-2048-257, which denoted that the sizes were 1799 ($257 * 7$, $\tau = 3$) for the input layer with a 7-frame context, 2048 for three hidden layers, and 257 for the output layer. The DNN-2 architecture was 1799-2048-2048-2048-514, with 514 ($257 * 2$) nodes at the output layer which was the only difference from DNN-1. The number of epochs 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 epochs was 50 and the mini-batch size T was set to 128. Input features of DNNs were globally normalized to zero mean and unit variance. Other parameter settings can be found in [59].

For the speech recognition system, the feature vector consisted 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 HMM with 32 Gaussian mixtures per state as specified in [50]. The chosen number of states for each word is the same as the setting in [50] and no extensive HMM retraining is performed. Please note that for SSC corpus, the ASR task has the limitation due to the constrained and simple grammar compared with other medium or large vocabulary recognition tasks. However, we intend to focus our attention mainly on speech separation part.

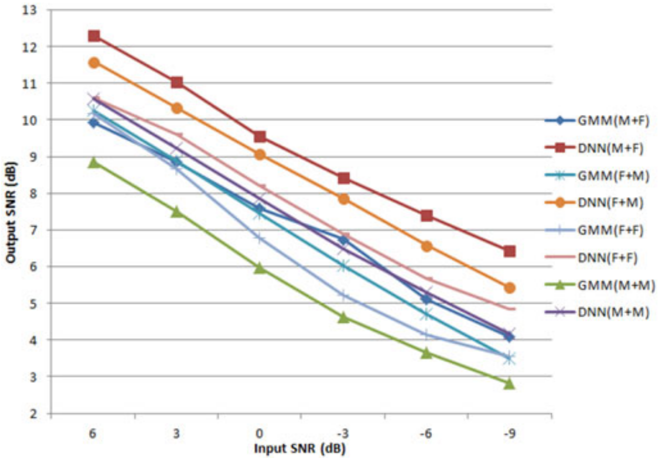


Fig. 6. Output SNR comparison of different approaches on the test set with four gender combinations in the 1 + 1 supervised mode.

A. *Separation Experiments on DNN-1*

In this section, the term “DNN” is referred to DNN-1.

1) 1+1 Supervised Speech Separation: In the 1 + 1 supervised mode, information of both the target and one interferer is provided in advance. Since training of each DNN combining one target and one interferer from the set of 34 speakers was time-consuming, a subset of 16 combinations of targets and interferers was randomly selected for training and evaluation. They were equally assigned for the four possible gender groupings, 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 for training the corresponding DNNs.

Fig. 6 gives an output SNR comparison of different separation approaches with the four gender combinations in the 1 + 1 supervised mode. Several observations can be made. First, all DNN systems significantly improved the output SNR over the GMM systems across different input SNRs and gender groupings. For example, about 2 dB enhancement was observed in the best case of M + F for almost all input SNR levels. On the other hand, for the worst F + F grouping, we could achieve at least an improvement of 1 dB for an input SNR level of 3 dB or less. 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 order of M + F, F + M, F + F, and M + M.

For comparing the objective intelligibility of synthesized speech, the corresponding STOI values are plotted in Fig. 7. Not surprisingly, DNN still consistently outperformed GMM

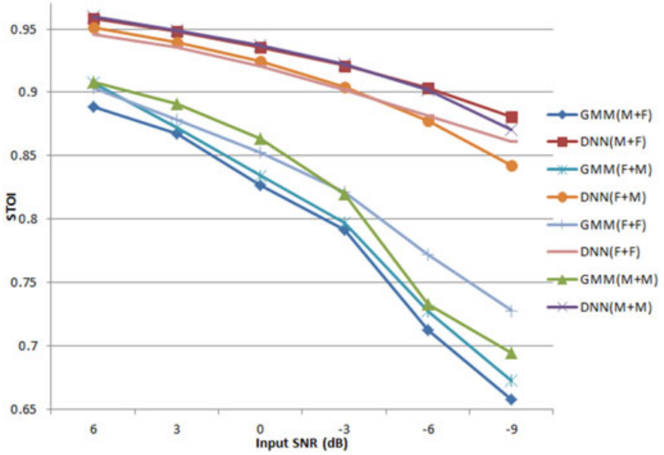


Fig. 7. STOI comparison of different approaches with four gender combinations in the 1 + 1 supervised mode on the test set.

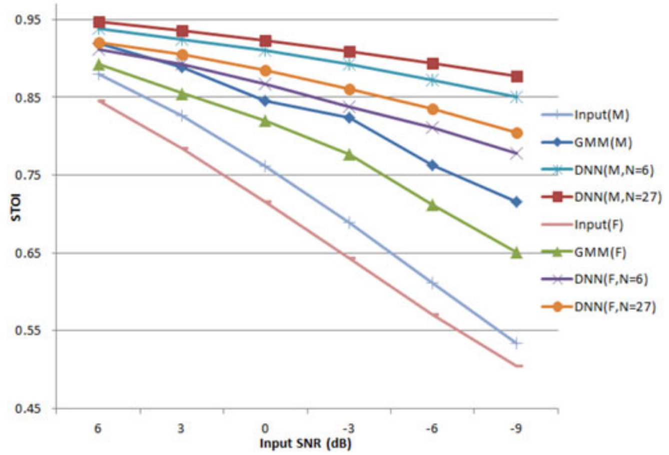


Fig. 8. STOI comparison of different approaches with the female (F) and the male (M) targets in the 1 + N supervised mode on the test set.

in all conditions. Moreover, it was interesting to note that our proposed DNN approach was more effective and robust than GMM, especially at low input SNR levels. Even at the -9 dB level, an STOI value of 0.87 for DNN with the M + M combination could still be achieved, which was better than that obtained by the corresponding GMM at about the 1 dB level, representing a very significant improvement of about 10 dB. Furthermore at least 0.9 STOI was obtained by the DNN approach at input SNR levels of -3 dB or above. It is also noted that the GMM approach produced the worst STOI in the easy M + F grouping than the other three gender combinations at all tested input SNR levels. It seems the DNN based results are more intuitive for gender grouping separation. On the other hand, the DNN approach managed to boost the STOI values to the best for the M + F combination among all gender groupings at almost all the input SNR levels which also agrees with our common belief that the different gender mixture is often easier to be handled than the same gender mixture.

2) 1 + N Speech Separation: In the 1 + N mode, 1 target and N interferers were used in the training stage to generate the mixed speech with two speakers. In the testing stage for separating the target, if the interfering speaker is one of the N interferers in the training stage, then it is still in a supervised mode. Otherwise, it is a semi-supervised mode with an unknown interferer. To test the effect of the number of mixing speakers, 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. This number of 27 is the most we can do with the SSC Corpus because there are only 34 speakers and the remaining speakers should be reserved as the unseen interferers in the separation stage. Training of DNNs with such an amount of data was time-consuming. So only one female target and one male target were randomly selected, and all the mixtures with those two targets on the test set were used for evaluation in the following experiments.

Fig. 8 shows an STOI comparison of different approaches with the female target (F) and the male target (M) in the 1 + N supervised mode. First of all, we found the STOI values for the male target speaker were always higher than those of the female target speaker in all testing conditions for input mixture, GMM and DNN systems with $N = 6$ and $N = 27$ at all six SNR levels, which might be due to the random selection of only one male and female target from the corpus. We also observed that by increasing N with more training data we could always improve STOI in the proposed DNN approach although the enhancement in objective intelligibility is relatively small. Similar to the results in Fig. 7, the STOI values of DNN were much better than those of GMM even with more confusing interferers included in training.

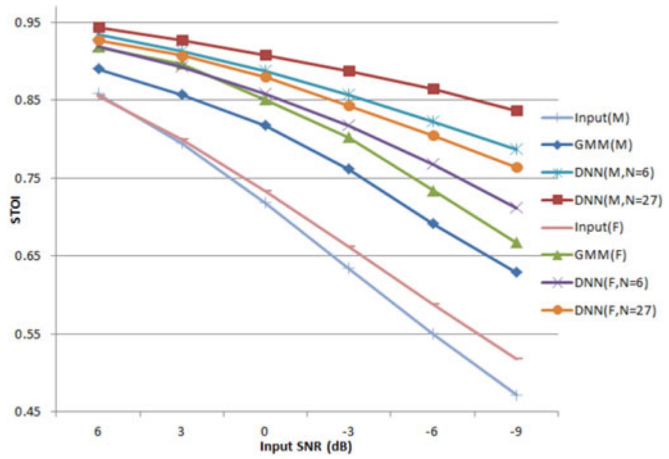


Fig. 9. STOI comparison of different approaches with the female target (F) and the male target (M) in the 1 + N semi-supervised mode on the test set.

One benefit of including various interfering speakers in DNN training is to have the trained DNNs perform speech separation in a semi-supervised mode. Fig. 9 lists an STOI comparison of the different approaches with the same female target (F) and the same male target (M) as in Fig. 8 for the 1 + N semi-supervised mode. Note that the results for GMM in Fig. 9 were still in a supervised mode. Similar observations as those in Fig. 8 could

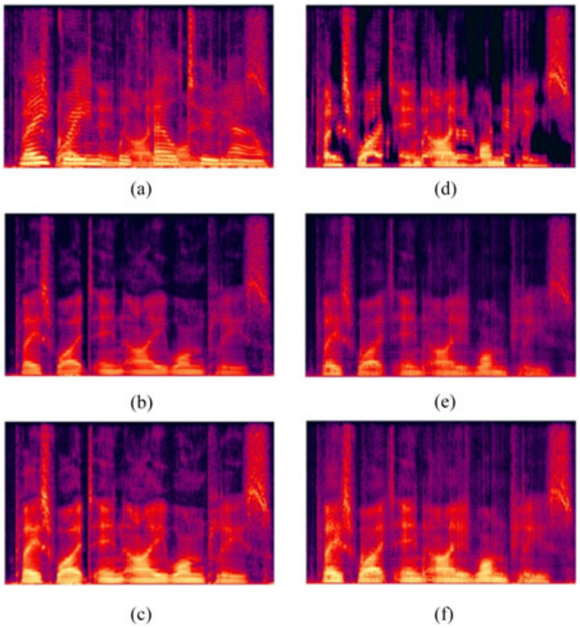


Fig. 10. Illustration of spectrograms for separating the target male utterance from the mixed utterance with a female interferer in the semi-supervised 1 + N mode ($N = 27$).

also be made although the STOI values in Fig. 9 tended to be slightly worse than those in the corresponding supervised system configuration shown in Fig. 8. There was only one exception that the semi-supervised DNN (F, $N = 6$) at 3 dB generated slightly worse STOI than supervised GMM at about 0.89 STOI. Overall, the DNN approach with $N = 27$ achieved consistently the best separation performance. These results were encouraging as our proposed DNN approach without any information about the interferers could outperform the conventional GMM approach with information of both the target and the interferer. This also 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.

Finally, the spectrograms of an example utterance are illustrated in Fig. 10 with Fig. 10(a) for a mixed utterance with a male target and a female interferer at −9 dB SNR and Fig. 10(b) for the original signal of the target male. Fig. 10(c) is a corresponding version with energy normalization (denoted as Target_N) as in [24] which is used as a reference for the spectrogram in Fig. 10(d) using a GMM approach where energy normalization should have been performed. Fig. 10(e) is the spectrogram of our proposed approach in the semi-supervised 1 + N mode ($N = 27$). To give a fair comparison with Fig. 10(d), the normalized version of our result (denoted as DNN_N) is also shown in Fig. 10(f). Clearly, our results were closer to the reference shown in Fig. 10(b) than that obtained with the GMM approach. Again it is also interesting to note that no interferer information was used in our DNN.

B. Separation Experiments on DNN-2

In this section, all the separation experiments are conducted on the 1 + N semi-supervised mode with $N = 10$. And for each target speaker, about 50 hours of mixed speech were used for DNN training. Fig. 11 shows a PESQ comparison of GMM, DNN-1 and DNN-2 approaches for 5 male (M) or 5 female (F) target speakers on the test set under different input SNRs. The PESQ performance was averaged across the target speakers with the same gender. The performances of DNN-2 were consistently better than those of both GMM and DNN-1 for all SNR levels, which confirmed that DNN-2 had a better generalization capacity over DNN-1. Furthermore, the PESQ performance of the male target for both DNN-1 and DNN-2 approaches 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 especially at low SNR levels.

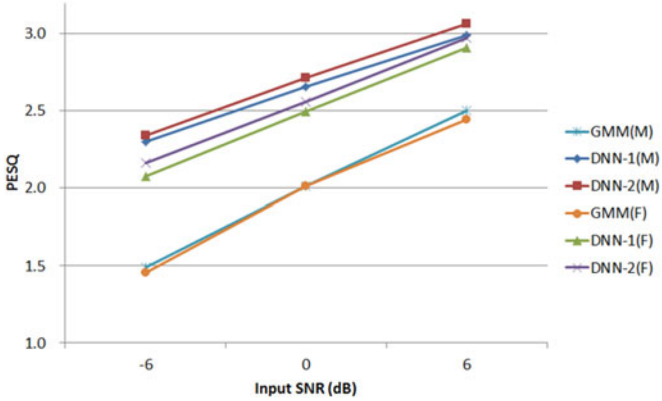


Fig. 11. PESQ comparison of GMM, DNN-1 and DNN-2 approaches for 5 male (M) or 5 female (F) target speakers on the test set under different input SNRs.

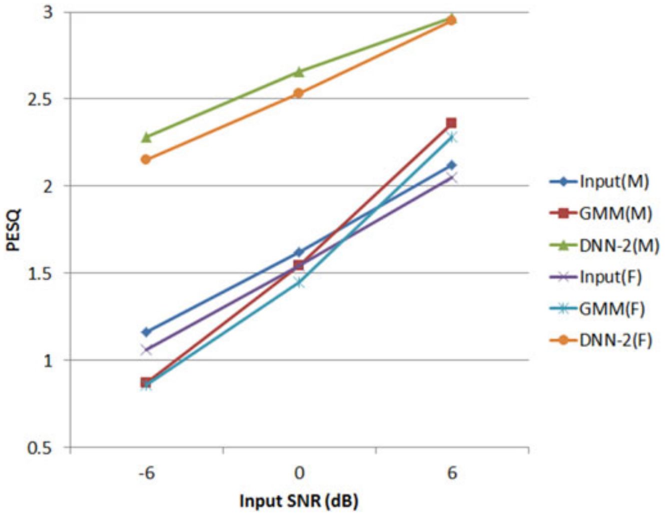


Fig. 12. PESQ comparison of input mixture, GMM and DNN-2 approaches for male (M) or female (F) interferers on the same data set as in Fig. 11.

Another benefit from DNN-2 is that the interfering speaker can also be separated. Fig. 12 lists a PESQ comparison of the original input mixture, GMM-enhanced and the DNN-2 approaches for male (M) or female (F) interferers on the same data set as in Fig. 11. The PESQ performance was averaged across the interferers with the same gender. First we could

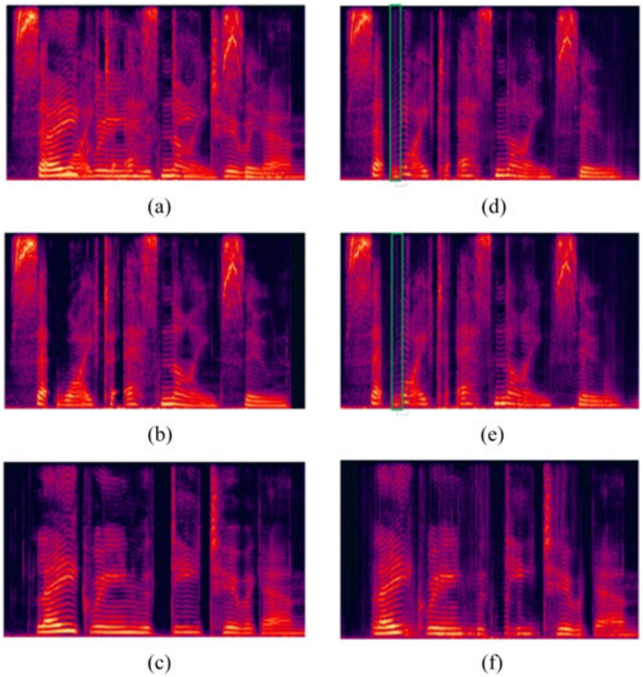


Fig. 13. Illustration of spectrograms for (a) input mixture with a female target and a female interferer at 0 dB SNR, (b) the female target, (c) the female interference, (d) DNN-1 separated female target, (e) DNN-2 separated female target, (f) DNN-2 separated female interferer.

observe that the GMM approach only outperformed the input mixture without separation at high SNR levels, e.g., 6 dB. Meanwhile, DNN-2 approach yielded very significant improvements over both the unprocessed input mixture and the supervised GMM approach which implied that the unseen interferers could also be well separated from the mixture. The performance gaps among different input SNRs of DNN-2 were much smaller than those in the GMM approach, which indicates that the DNN-2 approach was more effective under lower SNRs. For example, the PESQ improvement from 0.87 to 2.28 was observed for the male interferers at SNR $= -6$dB while the increment was from 2.36 to 2.97 at SNR $= 6$ dB. More interestingly, by a comparison of the PESQ performance in Figs. 11 and 12 with the same gender using the DNN-2 approach, there was only a small gap between the target speakers and the interferers. By considering that no information was provided for the unseen interferer, those results were quite encouraging and further confirmed the powerful predicting capability of DNN-based source separation.

Finally, the spectrograms of an utterance example are illustrated in Fig. 13. Fig. 13(a) is the spectrogram of mixed utterance with a female target and a female interferer at 0 dB. Fig. 13(b) is the spectrogram of the female target while Fig. 13(c) corresponds to the female interferer. Fig. 13(d) is the spectrogram of DNN-1 separated female target. Fig. 13(e) and (f) are the spectrograms of DNN-2 separated female target and female interference, respectively. For speech separation of the target speaker, DNN-2 generated good separation results which were close to the reference, and also outperformed DNN-1 , e.g., a better interference removal in the green rectangle regions shown in the beginning parts of the spectrograms in Fig. 13(d) and (e). Another interesting observation was that although there was no information about the interferer, we could still obtain a relatively good separation result of the unseen interferer in Fig. 13(f), especially in this confusing combination case of two female speakers, which further confirmed that our proposed DNN was effective in predicting unseen interferers by using multiple interfering speakers in training. Furthermore, DNN-2 demonstrated the potential of separating the target speaker from the interferer even the mixed speech signals were corrupted with background noises which was quite common in many real applications [60].

Table I gives the computational complexity of GMM, DNN-1 and DNN-2 approaches. Our experiments were conducted on a machine using Intel Xeon E5-2670 CPU with a clock rate of 2600 MHz. The training of DNN-1 and DNN-2 was accelerated by the Tesla K20 GPU. Obviously, both the model size and training time of DNN-1 and DNN-2 were comparable which were significantly larger than those in GMM model. This implies that the discriminative model (DNN) for the separation can make better use of model size to accommodate the scalable training data than the generative model (GMM). Meanwhile, the more computational complexity of DNN-based approach can also yield much better performance than GMM-based approach.

Based on the above experimental analysis, the DNN-2 architecture is used for all the following experiments in Sections V-C, V-D and V-E.

C. *Comparison With GSN Approach*

To further demonstrate the effectiveness of our DNN approach, we design a set of experiments to make a comparison with the GSN approaches [43], [44]. To conduct a fair comparison, we use the same data configurations on the GRID corpus as in [44] which is listed in Table II. The experiments on both supervised and semi-supervised (SD and SI in [44]) modes were investigated. Two female and male target speakers were selected for all tasks. In the supervised experiment, we totally trained 12 models with 3 models (corresponding to the remained 3 interferers) for each target speaker. For each model, the amount of training utterances were 2400 (randomly selected 400 utterances for each SNR level) while 50 testing utterances were used for each SNR level. In the semi-supervised experiment, 4 models were trained with each model using all interferers and mixing SNRs.

First, the PESQ performance of our DNN approach with different configurations is listed in Table III. As for the number of input frames as the acoustic context, 7 was a best choice across all SNRs in our experiments. And the DNN system with 3 hidden layers outperformed that with 2 hidden layers. In the following experiments of this section, the best configuration (7, 3) will be used by default. In the GSN approach [44], the setup to achieve the best performance was similar to our configuration (7, 2) in terms of model size.

Tables IV and V show the PESQ comparison of the DNN and GSN approaches for the semi-supervised (SI) and supervised (SD) tasks, respectively. As for the GSN approach, the results in [43], [44] were cited. The reason why our baseline

TABLE I
THE COMPUTATIONAL COMPLEXITY OF DIFFERENT APPROACHES FOR EACH SPEAKER MODEL

	Model configuration	Size (in MB)	Training data (in hour)	Training time (in hour)
GMM	128 dimensions, 256 mixtures	0.25	0.2	0.1
DNN-1	$1799(257 * 7)$-2048-2048-2048-257	48	90	48
DNN-2	$1799(257 * 7)$-2048-2048-2048-$514(257 * 2)$	50	90	48

TABLE II
THE SAME DATA CONFIGURATIONS AS IN [43], [44]

			# of utterances per case			
Task	Target IDs	Interferer IDs	Train	Validation	Test	Mixing SNRs (dB)
Supervised(SD)	1, 2, 18, 20	1, 2, 18, 20	400	50	50	−6, −3, 0, 3, 6, 9
Semi-supervised(SI)	1, 2, 18, 20	3, 4, 5, 6, 7, 8, 9, 10, 11, 15	50	5	5	−6,−3,0,3,6,9

TABLE III
THE PESQ COMPARISON OF THE DNN APPROACHES WITH DIFFERENT CONFIGURATIONS (N_{F} IS THE NUMBER OF INPUT FRAMES AND N_{L} IS THE NUMBER OF HIDDEN LAYERS) FOR THE SEMI-SUPERVISED (SI) TASK

$(N_{\mathrm{F}}, N_{\mathrm{L}})$	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB
(5, 3)	1.67	1.90	2.13	2.35	2.55	2.73
(7, 3)	1.76	1.99	2.21	2.43	2.63	2.81
(9, 3)	1.66	1.89	2.12	2.35	2.55	2.74
(7, 2)	1.73	1.94	2.15	2.35	2.53	2.70

TABLE IV
THE PESQ COMPARISON OF THE DNN AND GSN APPROACHES FOR THE SEMI-SUPERVISED (SI) TASK

	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB
The PESQ results in [43], [44]						
Baseline	1.37	1.65	1.81	2.07	2.38	2.59
GSN	1.62	1.87	2.06	2.29	2.55	2.75
Our results						
Baseline	1.34	1.56	1.80	2.03	2.26	2.47
DNN	1.76	1.99	2.21	2.43	2.63	2.81
DNN(20-fold)	2.05	2.28	2.51	2.73	2.93	3.10

TABLE V
THE PESQ COMPARISON OF THE DNN AND GSN APPROACHES FOR THE SUPERVISED (SD) TASK

	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB
The PESQ results in [43], [44]						
Baseline	1.60	1.85	2.08	2.32	2.56	2.77
GSN	2.09	2.30	2.53	2.75	2.94	3.14
Our results						
Baseline	1.55	1.84	2.11	2.33	2.52	2.68
DNN	2.51	2.69	2.84	2.99	3.13	3.26
DNN(20-fold)	2.81	2.97	3.12	3.28	3.41	3.53

TABLE VI
THE PESQ COMPARISON OF THE DNN APPROACHES USING 60-FOLD TRAINING DATA BY INCREASING THE MODEL COMPLEXITY (N_{U} IS THE NUMBER OF HIDDEN UNITS AND N_{L} IS THE NUMBER OF HIDDEN LAYERS)

$(N_{\mathrm{U}}, N_{\mathrm{L}})$	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB
(2048, 3)	2.48	2.67	2.84	2.99	3.12	3.24
(2048, 4)	2.41	2.60	2.77	2.92	3.05	3.17
(3072, 3)	2.40	2.59	2.75	2.9	3.02	3.14

performance was slightly different from that in [43], [44] was due to the randomly selected training utterances. Based on those results, several observations could be made. First, our DNN approach could significantly outperform GSN approach using the best configurations across all the SNRs for both SI and SD tasks. Second, even using the configuration (7, 2) as in Table III with the similar model size to GSN but not optimal, the corresponding results were still better than those of GSN in the SI task, especially for the low SNR cases. Finally, one of the most important issues, which was not emphasized by the previous work in [41], [44], was the increased amount of training data could bring very significant improvements of separation performance. DNN(20-fold) which used 20-fold synthesized training data compared with DNN and GSN was a good demonstration.

Finally, we investigate whether the separation performance can be further improved by using more parameters with more hidden units and layers in one single DNN. Based on Table VI, it was observed that more than 2048 hidden units or 3 hidden layers could not boost the system performances. The main reason might be the regression structure of DNN and the limitation of the SGD algorithm (at local optima).

D. Separation Experiments on SND-DNNs

1) SNR Estimation with a Single General DNN: The separation performance using SND-DNNs depends highly on how accurate the estimated SNR values of the mixture utterances are. However, the SNR estimation is a well studied problem [61], [62]. In this work, we adopt the Eq. (4) to estimate the SNR.

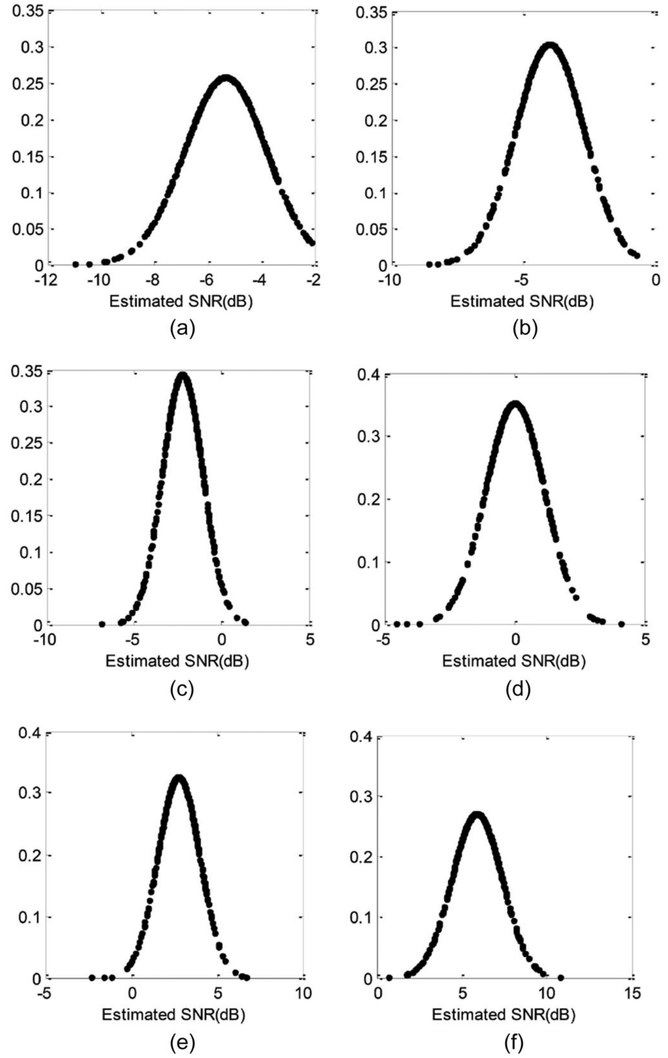


Fig. 14. Distributions of estimated SNR of input mixtures with various SNRs.

Fig. 14 shows the distributions of the estimated SNRs of the test data with different input SNRs. Several observations can be made. First, for all the testing cases except at an input SNR of 0 dB, our SNR estimation based on the separation results of the general DNN could give accurate decisions on most positive SNR and negative SNR cases. As for the 0 dB cases, there was no significant influence in the final decision because the 0 dB data set was included for training both positive and negative DNNs. Second, all distributions in Fig. 14 were unimodal. When the input SNR was above -3 dB, the distributions were centered exactly at the same input SNR values which indicated that a good SNR estimation was obtained by our approach. But for input SNR below -3 dB, e.g., in Fig. 14(a) and (b), the separation performance was degraded which led to the center shift to the right indicating over-estimated SNR values. Furthermore larger variances in the two distributions were obtained when compared with the other four higher SNR situations. Nonetheless, our proposed SNR estimation approach was accurate enough to make correct subsequent decisions because only one of the two SND-DNNs needed to be chosen.

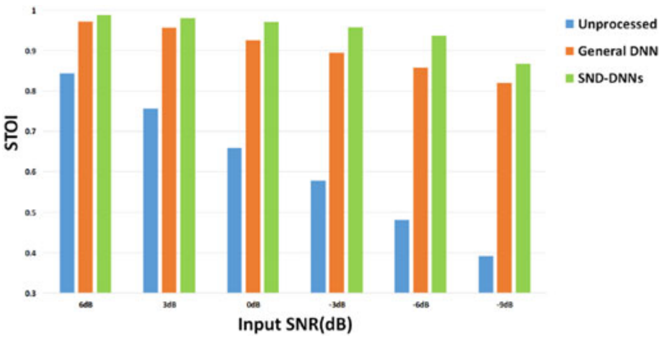


Fig. 15. STOI comparison of different approaches averaged across all 34 testing target speakers.

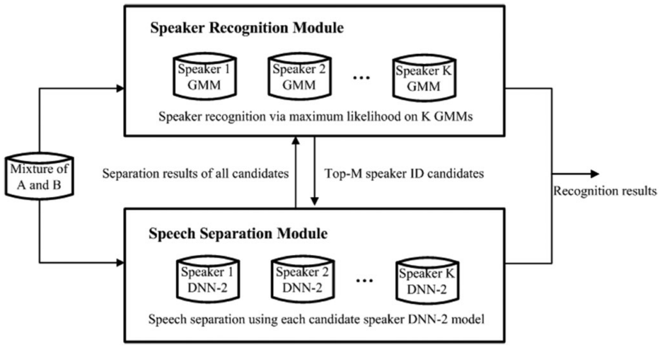

Speaker Recognition Module
Speaker 1 GMM
Speaker 2 GMM
...
Speaker K GMM
Speaker recognition via maximum likelihood on K GMMs
Mixture of A and B
Separation results of all candidates
Top-M speaker ID candidates
Recognition results
Speech Separation Module
Speaker 1 DNN-2
Speaker 2 DNN-2
...
Speaker K DNN-2
Speech separation using each candidate speaker DNN-2 model

Fig. 16. Speaker recognition module designed for separation and ASR.

2) Separation Performance With Two SND-DNNs: Fig. 15 lists an 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 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 and the lower the input SNR level the more the STOI improvement was observed, except in the difficult case when the input SNR level was at -9 dB. For example at SNR $= -6$ dB, the STOI was improved from 0.86 to 0.92, while the STOI was only improved from 0.96 to 0.98 when the input SNR was at 3 dB.

E. ASR Experiments After Speech Separation

Finally, the effectiveness of the proposed DNN based separation approach as an acoustic pre-processing module is further verified for robust speech recognition in SSC [49]. In [51], a speaker recognition module with more than 98% accuracy was implemented. However, our approach is originally designed for the scenarios which aim to separate a target speaker (already known) from the unknown mixture, so the speaker recognition is not necessary. But as a comprehensive study in general, we still propose a novel speaker recognition algorithm in Fig. 16 which is described as:

TABLE VII
THE RECOGNITION ACCURACY (IN %) OF OUR PROPOSED MULTI-PASS SPEAKER RECOGNITION ALGORITHM ON THE TEST SET

	6 dB	3 dB	0 dB	−3 dB	−6 dB	−9 dB
A or B (first-pass)	99.83	99.50	99.17	99.67	99.67	100
A or B (second-pass)	100	100	100	100	100	100
A and B (third-pass)	100	100	100	100	100	100

Step 1: Building models of all K speakers for both speaker recognition and speech separation modules
For the speaker recognition, use 39-dimensional perceptual linear prediction (PLP) features of all speakers to train a GMM as the universal background model [64], which is then adopted to initialize speaker dependent GMMs via MAP estimation [63]. For the speech separation module, the DNN-2 models for all speakers as the targets are built.

Step 2: First-pass recognition
For the input mixture with speaker A and B, use the GMM systems in Step1 to perform the first-pass recognition with the generated top-M ($M = 4$) speaker ID candidates for A or B fed to the separation module.

Step 3: First-pass separation
For each speaker ID candidate, the speech waveform of target output is separated from the input mixture using the corresponding DNN-2 model.

Step 4: Second-pass recognition
With the M separated waveforms in Step3 as the candidates, the second-pass recognition is conducted to select the top-1 speaker ID for A or B.

Step 5: Second-pass separation
Using the DNN-2 model of the speaker ID provided by Step4, we conduct the second-pass separation for the input mixture to obtain the dual outputs, namely the target speech and interferer speech.

Step 6: Third-pass recognition
With the dual outputs in Step5, the third-pass recognition can finally identify both speaker A and B in the mixture.

Through this procedure, the speaker recognition results of the multi-pass recognition can be achieved as in Table VII. Obviously, after the first-pass recognition (Step2), the recognition accuracy of one speaker (A or B) was not perfect, especially at 0 dB as this was the most confusing case. But the top-4 results can guarantee that at least one speaker of the mixture can be perfectly detected. Then the second-pass recognition (Step4) can accurately identify one speaker (A or B). Finally both A and B can be perfectly recognized in the third-pass recognition. So this multi-pass algorithm with the collaboration of speaker recognition and speech separation is superior to that in [51].

TABLE VIII
THE PERFORMANCE (WORD ACCURACY IN %) COMPARISON OF THE BASELINE, THE GENERAL DNN AND THE SND-DNNS APPROACH, AVERAGED ACROSS THE MIXTURE DATA WITH THE SAME GENDER AND DIFFERENT GENDER OF THE TEST SET

	6 dB	3 dB	0 dB	−3 dB	−6 dB	−9 dB	Avg.
16 kHz waveform, SD acoustic models							
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
25 kHz waveform, SI acoustic models							
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
25 kHz waveform, SD(SI + MAP) acoustic models							
Baseline	66.9	51.2	36.8	24.1	15.4	10.2	34.1
SND-DNNs	**95.8**	**94.4**	**93.7**	**91.7**	**88.5**	**83.6**	**91.3**
The best results in [50]							
SSC	93.0	92.5	91.5	89.5	87.0	79.0	88.8

In Table VIII, we report a performance (word accuracy in %) comparison of the baseline, the general DNN, and the SND-DNNs averaged across the mixture data of the test set at six SNR levels, ranging from −9 dB to +6 dB, for the standard 16 kHz waveforms. To make a fair comparison with the competition results using original waveforms, we also tested the recognition performance of our proposed SND-DNNs for the 25 kHz waveforms. Note that for ASR acoustic models used in the first two baseline systems, speaker-dependent models directly trained from the data of each speaker (SD) were used in the 16 kHz cases while speaker-independent (SI) models were adopted in the 25 kHz case with no retraining, i.e., the GMM-HMMs for ASR were officially provided by SSC to demonstrate the effectiveness of our proposed separation based acoustic pre-processing. Furthermore, another type of speaker-dependent models by using MAP adaptation on the provided speaker-independent models for each speaker [63] (SI + MAP) was also tested for the 25 kHz cases. For 25 kHz cases, the frame length and shift for DNN-based speech separation were set to 20 msec and 10 msec, respectively. This configuration differed from the setting of 32 msec and 16 msec in 16 kHz cases, respectively. 512-point DFT was used by zero padding to the 500 samples in one frame. Finally the 257-dimensional log-power spectral features were generated for DNN-based separation in 25 kHz cases with the same dimension as in 16 kHz cases.

The general DNN achieved 82.6% on the average over six SNR levels, representing significant performance improvements over the baseline system which was only at 23.0% without speech separation. It is noted that the accuracy rate increases were observed at all SNR levels, showing a good speech separation indeed alleviates some difficulty in dealing with residual noise after separation. On top of the general DNN, SND-DNNs consistently yielded additional performance gains for all testing cases, especially at low SNRs, e.g., at −3 dB, a relative word error rate (WER) reduction of 33.7% (from a WER of 18.7% to 12.4%) was observed. In average, the WER was decreased from 17.4% to 13.1%, with an absolute 4.3% WER reduction.

Furthermore using the 25 kHz waveforms, our proposed SND-DNNs approach with a single set of speaker-independent GMM-HMMs, under all SNRs, consistently outperformed the best results in the competition [50] (the average recognition accuracy of the same gender and different gender cases without the same talker cases). For example, relative WER reductions of 27.1% and 13.8% were observed at SNR levels of 6 dB and −9 dB, respectively. Moreover an overall relative WER reduction of 11.6% averaged across the whole test set was achieved. By considering that the best system [50] used both speech separation in the front-end and a complicated joint decoding framework of the target and interferer in the back-end with multiple sets of SD GMM-HMMs [51], our purely front-end approach based on SND-DNNs with a single set of SI GMM-HMMs is quite effective in terms of recognition accuracy, efficiency, and model compactness. And we expect additional post-processing could further increase the word accuracy. Finally, the improved SI + MAP GMM-HMMs with the SND-DNNs approach on the 25 kHz waveforms gave the best recognition performance, with the overall relative WER reductions of 12.1% and 22.3% over the SND-DNNs approach with SI GMM-HMMs and the best system in [50], respectively.

To further examine the significance of improvements over the other conventional approaches in our separation and recognition experiments, here we adopt a "matched pair test" in [65], [66] for significance test, which is a two-tailed test with the null hypothesis that there is no performance difference between the two systems. We use a minimum value of p to indicate a significance difference at the level of p in the statistical significance tests. We found that p value was always less than 0.001 for all cases indicating the improvement significance of our approach.

VI. Conclusion and Discussions

We have proposed a novel framework of speech separation based on DNN. The effectiveness is demonstrated in both 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. With the additional requirements of predicting the speech feature of the interesting speaker we believe the DNN architecture with dual outputs is more powerful than the architecture with the single output. 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 algorithms and applications. To achieve high model resolutions, two SND-DNNs, namely positive and negative DNNs, demonstrate to 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 best system in the competition 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.

Due to the limited availability of the 25 kHz waveforms in the SSC corpus, we were able to conduct ASR experiments at a higher sampling rates than the conventionally-adopted 16 kHz rates commonly used in the speech community for almost 50 years. It seems a high sampling rate for speech might be needed when we research into new challenging problems, such as speech separation in low SNRs, speech separation of same gender mixing, speech de-reverberation, and automatic speech and speaker recognition of distorted signals. We believe our encouraging ASR results for the 25 kHz speech signals could bring out an awareness for the research communities to again address signal processing issues that was very active in the 1960s and 1970s.

References

[1] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 434–444, Feb. 1997.

[2] J.-F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Process.*, vol. 44, no. 12, pp. 3017–3030, Dec. 1996.

[3] J.-F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal Process. Lett.*, vol. 4, no. 4, pp. 112–114, Apr. 1997.

[4] S. Makino, T. W. Lee, and H. Sawada, *Blind Speech Separation*. New York, NY, USA: Springer-Verlag, 2007.

[5] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 1–12, Jan. 2007.

[6] M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2002, pp. 881–884.

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

[8] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft masking filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2299–2310, Nov. 2007.

[9] M. Stark, M. Wohlmayr, and F. Pernkopf, "Source-filter-based single-channel speech separation using pitch information," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 242–255, Feb. 2011.

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

[11] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1766–1776, Aug. 2007.

[12] D. L. Wang and G. J. Brown, *Computational, Auditory Scene Analysis: Principles, Algorithms and Applications*. Hoboken, NJ, USA: Wiley, 2006.

[13] 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, May 1999.

[14] M. Wu, D. L. Wang, and G. J. Brown, " A multi-pitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Process.* vol. 11, no. 3, pp. 229–241, May 2003.

[15] Y. Shao and D. L. Wang, "Model-based sequential organization in cochannel speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 289–298, Jan. 2006.

[16] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2067–2079, Nov. 2010.

[17] K. Hu and D. L. Wang, "An unsupervised approach to cochannel speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 1, pp. 120–129, Jul. 2013.

[18] R. E. Yantorno, "Co-channel speech and speaker identification study," Final Report for Summer Research Faculty Program, Air Force Office of Scientific Research, Speech Processing Lab, Rome Labs, New York, NY, USA, 1998.

[19] R. E. Yantorno, "Co-channel speech study," Final Report for Summer Research Faculty Program, Research Laboratory AFRL/IF, Speech Processing Lab, Rome Labs, New York, NY, USA, 1999.

[20] J. Ming, R. Srinivasan, D. Crookes, and A. Jafari, “CLOSE—A data-driven approach to speech separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1355–1368, Jul. 2013.

[21] R. Weiss and D. Ellis, “Speech separation using speaker-adapted eigenvoice speech models,” *Comput. Speech Lang.*, vol. 24, pp. 16–29, 2010.

[22] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[23] 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, Nov. 2010.

[24] K. Hu and D. L. Wang, “An iterative model-based approach to cochannel speech separation,” *EURASIP J. Audio, Speech, Music Process.*, vol. 14, pp. 1–11, 2013.

[25] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[26] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Process.*, vol. 10, no. 1, pp. 19–41, 2000.

[27] 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.

[28] G. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large vocabulary speech recognition,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[29] A. Mohamed, G. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[30] G. Hinton *et al.* “Deep neural networks for acoustic modeling in speech recognition,” *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[31] Y. X. Wang and D. L. Wang, “Towards scaling up classification-based speech separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

[32] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “An experimental study on speech enhancement based on deep neural networks,” *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[33] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7962–7966.

[34] S. Kang, X. Qian, and H. Meng, “Multi-distribution deep belief network for speech synthesis,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 8012–8016.

[35] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[36] 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,” in *Proc. Int. Conf. Signal Process.*, 2014, pp. 473–477.

[37] J. T. Chen, Y. X. Wang, and D. L. Wang, “Noise perturbation improves supervised speech separation,” in *Latent Variable Analysis and Signal Separation*. New York, NY, USA: Springer-Verlag, 2015, pp. 83–90.

[38] F. Weninger, F. Eyben, and B. Schuller, “Single-channel speech separation with memory-enhanced recurrent neural networks,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 3737–3741.

[39] J. Le Roux, J. R. Hershey, and F. Weninger, “Deep NMF for speech separation,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 66–70.

[40] P.-S. Huang, M. Kim, M. H. Johnson, and P. Smaragdis, “Deep learning for monaural speech separation,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 1581–1585.

[41] P.-S. Huang, M. Kim, M. H. Johnson, and P. Smaragdis, “Joint optimization of masks and deep recurrent neural networks for monaural source separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.

[42] M. Zöhrer and F. Pernkopf, “Single channel source separation with general stochastic networks,” in *Proc. INTERSPEECH*, 2014, pp. 978–982.

[43] M. Zöhrer and F. Pernkopf, “Representation models in single channel source separation,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 713–717.

[44] M. Zöhrer, R. Peharz, and F. Pernkopf, “Representation learning for single-channel source separation and bandwidth extension,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2398–2409, Dec. 2015.

[45] 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,” in *Proc. Int. Symp. Chin. Spoken Lang. Process.*, 2014, pp. 250–254.

[46] F. Xie and D. V. Compernolle, “A family of MLP based nonlinear spectral estimators for noise reduction,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1994, pp. 53–56.

[47] Y.-H. Tu, J. Du, L.-R. Dai, and C.-H. Lee, “Speech separation based on signal-noise-dependent deep neural networks for robust speech recognition,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 61–65.

[48] Y.-H. Tu, J. Du, Y. Xu, L.-R. Dai, and C.-H. Lee, “Deep neural network based speech separation for robust speech recognition,” in *Proc. Int. Conf. Signal Process.*, 2014, pp. 532–536.

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

[50] M. Cooke, J. R. Hershey, and S. J. Rennie, “Monaural speech separation and recognition challenge,” *Comput. Speech Lang.*, vol. 24, no. 1, pp. 1–15, 2010.

[51] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, “Super-human multi-talker speech recognition: A graphical modeling approach,” *Comput. Speech Lang.*, vol. 24, no. 1, pp. 44–66, 2010.

[52] C. Weng, D. Yu, M. Seltzer, and J. Droppo, “Single-channel mixed speech recognition using deep neural networks,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 5669–5673.

[53] Y. Bengio, “Learning deep architectures for AI,” *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[54] G. Hinton, S. Osindero, and Y. Teh, “A fast learning algorithm for deep belief nets,” *Neural Comput.*, vol. 18, pp. 1527–1554, 2006.

[55] X. Zhao, Y. Shao, and D. L. Wang, “CASA-based robust speaker identification,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1608–1616, Jul. 2012.

[56] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, 2006.

[57] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4214–4217.

[58] *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, ITU-T Rec. P.862, 2001.

[59] G. Hinton, “A practical guide to training restricted Boltzmann machines,” University of Toronto, Toronto, ON, Canada, UTML TR 2010-003, 2010.

[60] T. Gao, J. Du, L. Xu, C. Liu, L.-R. Dai, and C.-H. Lee, “A unified speaker-dependent speech separation and enhancement system based on deep neural networks,” in *Proc. China Summit Int. Conf. Signal Inform. Process.*, 2015, pp. 687–691.

[61] A. Narayanan and D. L. Wang, “A CASA-based system for long-term SNR estimation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 9, pp. 2518–2527, Nov. 2012.

[62] P. Papadopoulos, A. Tsiartas, J. Gibson, and S. Narayanan, “A supervised signal-to-noise estimation of speech signals,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 8237–8241.

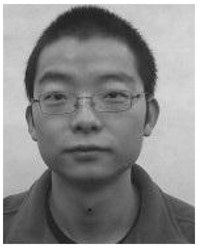
[63] J. L. Gauvain and C. H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observation of Markov chains,” *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[64] D. A. Reynolds, T. F. Quatieti, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[65] L. Gillick and S. J. Cox, “Some statistical issues in the comparison of speech recognition algorithms,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1989, pp. 532–535.

[66] D. S. Pallet, W. M. Fisher, and J. G. Fiscus, “Tools for the analysis of benchmark speech recognition tests,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1990, pp. 97–100.

Jun Du received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively. From 2004 to 2009, he was with iFlytek Speech Lab of USTC. During the above period, he worked as an Intern twice for nine months at Microsoft Research Asia (MSRA), Beijing, China. In 2007, he also worked as a Research Assistant for six months in the Department of Computer Science, The University of Hong Kong. From July 2009 to June 2010, he worked with iFlytek Research on speech recognition. From July 2010 to January 2013, he joined MSRA as an Associate Researcher, working on handwriting recognition, OCR, and speech recognition. Since February 2013, he has been with the National Engineering Laboratory for Speech and Language Information Processing of USTC.



Yanhui Tu received the B.S. degree from the Department of Electronic Information Engineering, Yunnan University, Yunnan, China, in 2013. He is currently working toward the Ph.D. degree at the University of Science and Technology of China, Hefei, China. His current research interests include speech separation and robust speech recognition.

Li-Rong Dai was born in China in 1962. He received the B.S. degree in electrical engineering from Xidian University, Xian, China, in 1983, and the M.S. degree from the Hefei University of Technology, Hefei, China, in 1986, and the Ph.D. degree in signal and information processing from the University of Science and Technology of China (USTC), Hefei, in 1997. He joined the University of Science and Technology of China in 1993. He is currently a Professor at the School of Information Science and Technology, USTC. His current research interests include speech synthesis, speaker and language recognition, speech recognition, digital signal processing, voice search technology, machine learning, and pattern recognition. He has published more than 50 papers in these areas.

Chin-Hui Lee is a Professor at the School of Electrical, and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Before joining academia in 2001, he had 20 years of industrial experience ending in Bell Laboratories, Murray Hill, NJ, USA, as a Distinguished Member of Technical Staff, and the Director of the Dialogue Systems Research Department. He is a Fellow of ISCA. He has published more than 400 papers, and 30 patents, and was highly cited for his original contributions with an h-index of 66. He received numerous awards, including the Bell Labs President's Gold Award in 1998, and the SPS's 2006 Technical Achievement Award for "Exceptional Contributions to the Field of Automatic Speech Recognition." In 2012, he was invited by ICASSP to give a plenary talk on the future of speech recognition. In the same year, he received the ISCA Medal in scientific achievement for pioneering and seminal contributions to the principles and practice of automatic speech and speaker recognition.