A Gender Mixture Detection Approach to Unsupervised Single-Channel Speech Separation Based on Deep Neural Networks

Yannan Wang, Jun Du, Li-Rong Dai, and Chin-Hui Lee, *Fellow, IEEE*

***Abstract*—We propose an unsupervised speech separation framework for mixtures of two unseen speakers in a single-channel setting based on deep neural networks (DNNs). We rely on a key assumption that two speakers could be well segregated if they are not too similar to each other. A dissimilarity measure between two speakers is first proposed to characterize the separation ability between competing speakers. We then show that speakers with the same or different genders can often be separated if two speaker clusters, with large enough distances between them, for each gender group could be established, resulting in four speaker clusters. Next, a DNN-based gender mixture detection algorithm is proposed to determine whether the two speakers in the mixture are females, males, or from different genders. This detector is based on a newly proposed DNN architecture with four outputs, two of them representing the female speaker clusters and the other two characterizing the male groups. Finally, we propose to construct three independent speech separation DNN systems, one for each of the female–female, male–male, and female–male mixture situations. Each DNN gives dual outputs, one representing the target speaker group and the other characterizing the interfering speaker cluster. Trained and tested on the speech separation challenge corpus, our experimental results indicate that the proposed DNN-based approach achieves large performance gains over the state-of-the-art unsupervised techniques without using any specific knowledge about the mixed target and interfering speakers being segregated.**

***Index Terms*—Unsupervised speech separation, speaker clustering, gender mixture detection, deep neural network, speaker dissimilarity measure.**

I. Introduction

CO-CHANNEL speech separation [1], referring to separating a speech component of interest from noisy speaker mixtures, has a variety of important applications, e.g., automatic speech recognition (ASR) [2] in the recent Speech Separation Challenge (SSC) [3]. We often formulate the problem with two mixing speakers as follows:

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

where $\boldsymbol{x}^{\mathrm{m}}$ is the mixed speech signal while $\boldsymbol{x}^{\mathrm{t}}$ and $\boldsymbol{x}^{\mathrm{i}}$ refer to speech of the target and interfering speakers, respectively. Model-based approaches are widely used for speech separation in a supervised mode which generally build speaker-dependent models assuming the identities of the target and interfering speakers are known. Many approaches to model the speakers have been investigated. For instance, Roweis [4] employs the factorial hidden Markov model (FHMM) to learn the information of a speaker and then separates the speech mixture through computing a mask function and refiltering. Another probabilistic model named as factorial-max vector quantization (MAXVQ) is introduced in [5]. Moreover, 2-D Viterbi algorithms and loopy belief approximation have been adopted to conduct the inference in FHMM based approaches [6]. The layered FHMM incorporating temporal and grammar dynamics [7] performs quite well in monaural speech separation and recognition challenge. The Gaussian mixture model (GMM) [8] is also used in [9], [10] via minimum mean-square error (MMSE) estimation to resynthesize the speech signals. An iterative GMM-based approach is proposed in [11] based on a maximum a posteriori (MAP) estimator to overcome possible mismatches between the training and test conditions. Another popular approach is non-negative matrix factorization (NMF) [12], [13] which decomposes the signal into sets of bases and weight matrices. Recently a non-negative back-propagation algorithm is proposed in [14] to build a deep network with non-negative parameters.

The aforementioned supervised methods could achieve a satisfactory performance. However, they are not always applicable to practical scenarios due to a lack of prior knowledge of speakers. On the other hand in an unsupervised mode, computational auditory scene analysis (CASA) [15] is widely adopted in co-channel speech separation tasks. It is inspired by the ability of human auditory perception to recover signals of interest from background distractions. For example, in [16] pitch and amplitude modulation are employed to obtain the voiced components of co-channel speech through grouping estimated pitches. In [17] onset/offset-based segmentation and model-based grouping are introduced to deal with unvoiced portions. Unsupervised clustering for sequential grouping is adopted to convert

Manuscript received August 18, 2016; revised January 23, 2017 and April 4, 2017; accepted April 21, 2017. Date of publication May 2, 2017; date of current version June 5, 2017. This work was supported in part by the National Natural Science Foundation of China under Grants 61671422 and U1613211, in part by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDB02070006, and in part by the National Key Research and Development Program of China under Grant 2016YFB1001300. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hirokazu Kameoka. *(Corresponding author: Jun Du.)*

Y. Wang, J. Du, 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: wyn314@mail.ustc.edu.cn; jundu@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.2017.2700540

simultaneous streams to two clusters in [18] by maximizing the ratio of between-cluster and within-cluster distances. In general, there are two main stages in CASA approaches: segmentation and grouping [15]. Segmentation decomposes mixed speech into time-frequency (T-F) segments assumed to derive from the corresponding sound source. A simultaneous and sequential grouping assembles the T-F segments to generate independent streams. Moreover, a data-driven approach is proven to be effective which matches each mixed speech segment against a composite training segment to separate the underlying clean speech segments in [19]. In addition, a Bayesian NMF model [20] is proposed to separate tonal and percussive components from an audio signal in an unsupervised manner without any prior training.

Nowadays a deep learning framework called deep clustering is also proposed in [21] to assign contrastive embedding vectors to each T-F region of the spectrogram in order to predict the segmentation label of the target spectrogram from the input mixtures. In some other recent work [22]–[24], deep learning techniques, namely deep neural network (DNN) or recurrent neural network (RNN), have been adopted to model the highly non-linear mapping relationship from mixed speech to the target and interfering signals in a supervised or semi-supervised mode. More recently, a deep ensemble method named multicontext networks is presented in [25] to leverage contextual information sufficiently. In this work, we extend the DNN regression framework [23] to unsupervised speech separation of two speakers based upon the assumption that the larger the distance between competing speakers the better the mixed speakers could be separated. Intuitively, there is a distinct discrepancy between two speakers based on different pronunciation mechanisms, such as vocal tracts, fundamental frequency contours, dynamic ranges and speaking styles. These factors result in dissimilarities between speakers which might be characterized by certain distance measures. By adopting the i-vector technique [26] to represent each speaker, the distances among speakers of different genders are first visualized with a large margin to ensure a good separation. Furthermore, for the speakers within the same gender, if two groups can be divided in terms of maximizing the group distance, the segregation of two speakers from each group is also possible.

Inspired by this, we first divide the speakers into female and male groups and then further cluster each group into two sub-groups (F1/F2 as two female groups and M1/M2 as two male groups), to handle all gender-mixing cases in co-channel speech separation, namely male-male, female-female and female-male mixtures. Accordingly, we propose a two-stage detection/separation framework in an unsupervised setting. In the first stage, we adopt a DNN with four outputs, each representing the corresponding four speaker groupings (F1/F2 and M1/M2), to detect the gender combination and select the corresponding separator. In the second stage, the DNN with dual outputs representing two speaker groups, choosing from three gender-combination dependent separators, is applied to obtain two segregated results.

In contrast to the CASA approach, the main advantages of our approach are: 1) CASA uses the locally designed features as the psychoacoustic cues for the mid-level representation and scene organization while our approach adopts the most informative log-power spectra (LPS) features as the global representation; 2) we use a deep model for the detection/separation corresponding to the segmentation/grouping in CASA. The experiments on the SSC corpus further confirm that the proposed method can achieve a significantly better separation performance than the state-of-the-art CASA approach.

The rest of the paper is organized as follows. A characterization of the speaker dissimilarity is presented in Section II. The proposed DNN framework with a gender mixture detector followed by a speech separator is described in Section III. Training and testing data setup which is critical by considering the limited 34 speakers of SSC corpus is highlighted in Section IV. All experimental results are analyzed in Section V. Finally we summarize our findings in Section VI.

II. Characterization of Speaker Distances

As we aim at adopting speaker-independent DNN models to perform single-channel speech separation of two unseen speakers, in principle the prior information of the speaker separability should be leveraged in an unsupervised setting. Accordingly, the preliminary experiments on speaker dissimilarity measures were designed in the following subsections, which are used as evidences of the propose framework in the Section III.

A. I-Vector Based Measures

We adopt the recently emerged i-vector based speaker representation [26], [27] to measure the speaker dissimilarity. The core idea of i-vector extraction is that the speaker-independent and channel-independent supervector $\boldsymbol{s}$ can be formulated as:

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w} \tag{2}$$

where $\boldsymbol{m}$ is the mean supervector of LPS features for universal background model (UBM) [26], $\boldsymbol{w}$ is a latent variable with a standard normal distribution and the low-rank matrix $\boldsymbol{T}$ referred as the total variability matrix contains both speaker and channel variabilities. An i-vector $\boldsymbol{v}$ for the speech utterances which represents the speaker and channel information in a low-dimensional space is then obtained as the maximum a posterior point estimate of the latent variable $\boldsymbol{w}$.

Suppose we have N $(N = 34)$ speakers, then a distance matrix $\boldsymbol{D}$ of $N \times N$ dimensions can be generated with each element d_{ij} representing the Euclidean distance between the i-vectors of the ith and jth speakers:

$$d_{ij} = \| \boldsymbol{v}_i - \boldsymbol{v}_j \|_2 \tag{3}$$

where $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are K-dimension $(K = 100)$ i-vectors of two speakers trained with the utterances of each speaker to be described in Section IV.

B. Visualization of Speaker Distances

To visualize the similarity between two individual objects (the speakers in this work) in a low-dimensional space, each object to be studied can be represented by a point and the points are

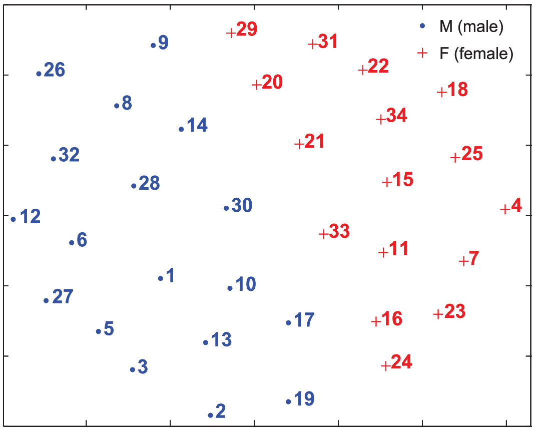

Fig. 1. MDS graph of the male and female speaker groups.

elaborately arranged in order to approximate the distances between pairs of objects. We use multidimensional scaling (MDS) [28], [29] to graphically describe the relationship conveyed by the aforementioned distance measures. MDS is a statistical technique to visualize the data through translating a table of dissimilarities between pairs of objects into a map where distances between the points match the dissimilarities as close as possible. Here the goal of MDS is to find an appropriate matrix $\boldsymbol{P}$ of $N \times L$ dimensions in which L is the target space dimension to be specified by the user ($L = 2$ in our experiments). And d_{ij}^{MDS}, the distance between the ith and jth row vectors of $\boldsymbol{P}$ ($\boldsymbol{p}_i$ and $\boldsymbol{p}_j$), should satisfy the condition:

$$d_{ij}^{\text{MDS}} = \| \boldsymbol{p}_i - \boldsymbol{p}_j \|_2 \approx d_{ij}, i, j = 1, 2, ..., N. \qquad (4)$$

Then we adopt a widely used nonlinear mapping criterion, namely the Sammon's mapping [30] to conduct MDS as follows:

$$\sigma_{\text{sammon}}^2(\boldsymbol{P}) = \frac{1}{\sum_{j<i} d_{ij}} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \frac{(d_{ij} - d_{ij}^{\text{MDS}})^2}{d_{ij}} \qquad (5)$$

where the distance errors are normalized by the distance in the original space.

The MDS graph for all 34 speakers of the SSC corpus is shown in Fig. 1. The blue-dot and red-cross marks represent the male and female speakers with the corresponding index numbers. We can observe that each speaker is surrounded by a group of neighboring speakers. Therefore a particular speaker could be characterized by such a group of neighboring speakers in an unsupervised manner. This figure also shows that the female and the male groups could be well separated with a large margin. Even within the same-gender group, the large distances between quite a few pairs of speakers demonstrate the feasibility of segregation. This is the motivation for the design of speaker groups in Section II-C and the proposed detection and separation framework in Section III. Furthermore, the cosine distance between i-vectors is also investigated. We can obtain similar visualization results to Fig. 1 and the same clustering results as discussed next in Section II-C.

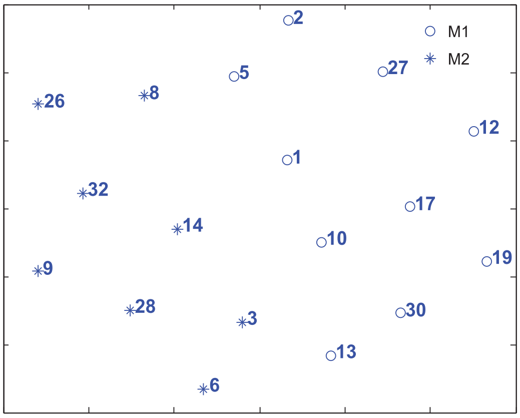

Fig. 2. MDS graph of male speaker groups after k-means clustering.

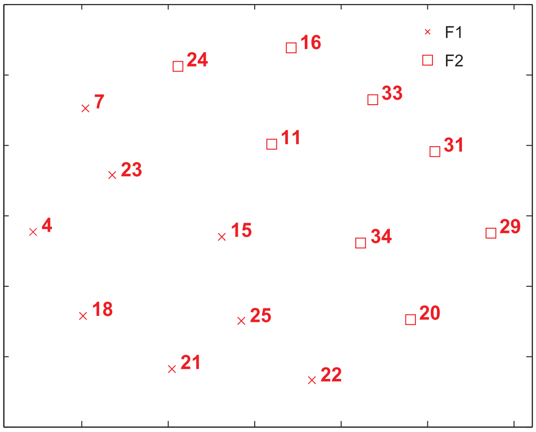

Fig. 3. MDS graph of female speaker groups after k-means clustering.

C. *Speaker Grouping*

In terms of the gender combination, a mixture of two speakers generally belongs to three cases, namely mixing of male-female (M-F), male-male (M-M), and female-female (F-F) speaker groups. So even the specific information of two unseen speakers could not be provided, unsupervised separation between them could still degenerate to the problem of conducting separation for two speaker groups if each group can well represent one mixing speaker. In Fig. 1, it is illustrated that the speakers could be well clustered by the different gender information, corresponding to the M-F mixture case. Interestingly, if k-means clustering [31] is applied to the MDS graph of 34 speakers, the generated clusters are exactly the same as the two gender groups in Fig. 1.

To handle the same-gender mixture cases (M-M or F-F), more speaker groups should be further designed, as illustrated in Figs. 2 and 3, redrawing the MDS graphs of the 18 male speakers and 16 female speakers, respectively. Clear margins between the two sub-groups (M1/M2 or F1/F2) within the same-gender group after applying k-means clustering can also be observed similar to that in Fig. 1. However, in terms of the separation difficulty, the same-gender mixtures could be further divided into two cases. The first case is that the two mixing speakers are from different sub-groups (denoted as M-M-D or F-F-D), e.g., one speaker from the M1 group and the other speaker from the M2 group. The second case is more challenging with the two

Algorithm 1: Procedure of speaker grouping based on k-means clustering.

```
Step 1: I-vector extraction
    Extracting the 100-dimension i-vector $\boldsymbol{v}$ which
  represents the speaker characteristics with the
  corresponding utterances of each speaker.
Step 2: Euclidean distance matrix generation
    Calculating the Euclidean distance matrix based on the
  i-vectors of all the speakers.
Step 3: MDS dimension reduction
    Conducting the MDS algorithm under the Sammon's
  mapping criterion.
Step 4: $k$-means clustering
    Running $k$-means clustering 50 times and picking the
  clustering results corresponding to the optimal objective
  value via the similar way as in [32] to alleviate local
  minima problems of $k$-means clustering.
```

TABLE I
THE AVERAGE DISTANCE BASED ON THE MATRIX $\boldsymbol{D}$ ACROSS ALL SPEAKER PAIRS FOR EACH OF 5 GENDER COMBINATIONS

Combination	M-F	M-M		F-F	
Distance	17.73	M-M-D	M-M-S	F-F-D	F-F-S
		17.27	16.36	16.55	15.65

mixing speakers from the same sub-groups (denoted as M-M-S or F-F-S). As mentioned above, the speaker grouping procedure can be summarized as follows:

Table I gives the averaged distances based on the matrix $\boldsymbol{D}$ across all speaker pairs for each of the five gender combinations corresponding to five input mixture cases. First, the M-F combination yields the largest distance which implies that the different-gender mixtures should have a better separability than the same-gender mixtures. Second, for the same-gender mixture, the case of the mixing speakers from different sub-groups (M-M-D or F-F-D) has a larger distance than the corresponding cases of the mixing speakers from the same sub-groups (M-M-S or F-F-S). Finally, the F-F combination seems to be more challenging to distinguish than the M-M combination. All these observations are in accordance with the criterion for setting those groups. More importantly, from the analysis of the subsequent experiments, the distance measure in Table I can well predict the difficulty of speech separation.

III. DNN-BASED UNSUPERVISED SPEECH SEPARATION

A. System Architecture

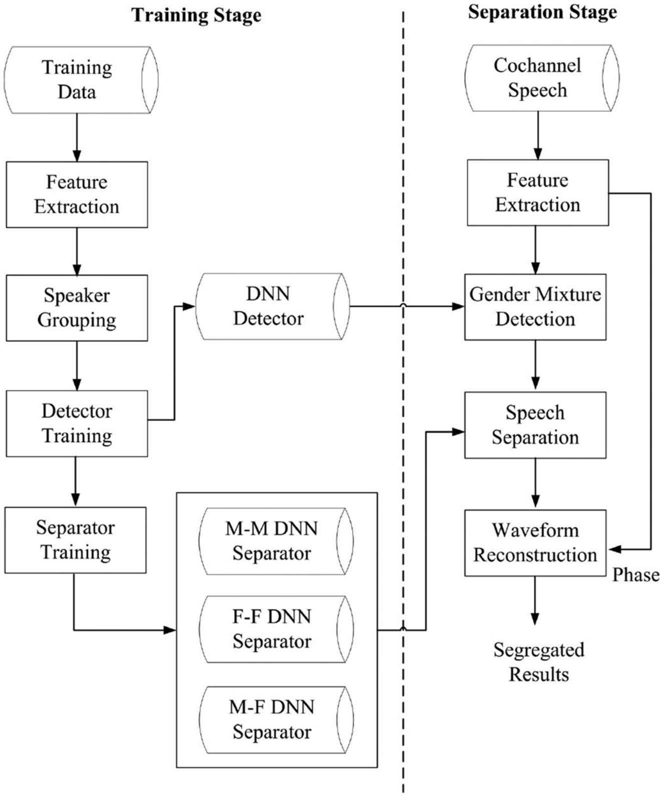

Fig. 4. The proposed unsupervised speech separation system.

Fig. 4 presents the proposed system architecture based on DNNs for unsupervised co-channel speech separation. We first construct four speaker clusters as described in Section II-C, denoted as M1, M2, F1 and F2, with the training speaker data of from each of the four groups. Then the gender mixture detector is implemented by a DNN with four outputs corresponding to the four corresponding speaker groups. Finally the speech mixtures of different combinations are adopted to train the set of DNN-based separators. Specifically, three DNN separators are designed, including the M-M, F-F and M-F separators, to cover all possible gender combinations. In the separation stage, after feature extraction, the mixed speech is first processed by a gender mixture detector to determine the type of gender combination. Then speech separation is conducted with the corresponding DNN separator obtained in the training stage. Compared with the conventional CASA system in [15], the outputs of the first several hidden layers in the DNN separator are similar to the mid-level representation of acoustic features while the segregation based on the last several layers of DNN is analogous to the scene organization. One advantage of the DNN-based approach is that the design of handcraft features as the psychoacoustic cues in CASA is not necessary as the deep feature representations can be automatically learned. In the following subsections, the DNN-based detector and separators are elaborated.

B. Gender Mixture Detection

To show the importance of the gender mixture detector and the effectiveness of the DNN-based approach, we first introduce a Gaussian mixture model - universal background model (GMM-UBM) method widely used in the speaker recognition community [33], [34] as a comparison in experiments. With a UBM for the alternative speaker representation and a form of Bayesian adaptation to derive the speaker models from the UBM, two GMMs representing male speakers and female speakers are trained and then used to determine the gender identities of mixed

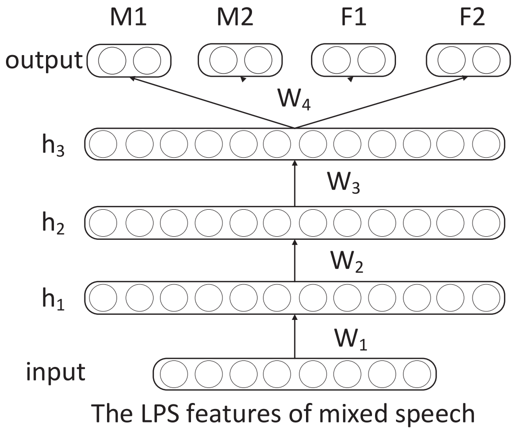

Fig. 5. A DNN-based frontend for the proposed gender mixture detector.

speech as follows:

$$l^{\mathrm{M}} - l^{\mathrm{F}} \begin{cases} > \alpha & \text{M-M} \\ < -\alpha & \text{F-F} \\ \in [-\alpha, \alpha] & \text{M-F} \end{cases} \tag{6}$$

where l^{M} and l^{F} are log-likelihoods of the speech mixture utterance given the male and female GMMs, respectively. α is a threshold to balance the accuracy among different gender combinations.

The GMM-UBM model often cannot well distinguish the gender information in mixed speech. Accordingly, we first adopt a classification DNN (CDNN) detector to make a decision on which combination (M-M, F-F, or M-F) the input speech utterance belongs to. The CDNN is a frame-level classifier with 4 softmax nodes representing silence, male speech (single-male or male-male mixed speech), female speech (single-female or female-female mixed speech), and male-female mixed speech, respectively. In the training stage, with the two mixing utterances with utterance-level gender information and frame-level speech/non-speech information via a simple energy-based approach, all the input frames of mixed speech can be assigned to one of these four classes. In the detection stage, the following rule is adopted for the decision of gender combination through comparing the probabilities of different gender combinations across all frames of current utterance as follows:

$$\begin{cases} \frac{1}{T}\sum_{t=1}^{T}\left(P_t^{\mathrm{M}} + 0.5 * P_t^{\mathrm{MF}}\right) > \lambda & \text{M-M} \\ \frac{1}{T}\sum_{t=1}^{T}\left(P_t^{\mathrm{F}} + 0.5 * P_t^{\mathrm{MF}}\right) > \lambda & \text{F-F} \\ \qquad\qquad\qquad\text{else} & \text{M-F} \end{cases} \tag{7}$$

where P_t^{M}, P_t^{F} and P_t^{MF} denote the softmax outputs of CDNN for tth frame, respectively. λ is a threshold closed to 0.64 which will be discussed in the experiments. The second item $0.5 * P_t^{\mathrm{MF}}$ in (7) can reduce the misclassifications of the same-gender combination utterances to M-F utterances.

Moreover we propose a regression DNN (RDNN) detector with 4 outputs to predict the gender mixture of the two mixing speakers. As shown in Fig. 5, the inputs are the LPS features of mixed speech with neighboring frames generated by speakers from the four different groups while the four outputs are the LPS features of the corresponding target and interfering speakers that belonging to two of the four groups, namely M1, M2, F1, and F2. In training, the DNN parameters are randomly initialized and then optimized by jointly minimizing the mean square errors between the DNN outputs and the target LPS features:

$$E_1 = \frac{1}{T}\sum_{t=1}^{T}\Big(\left\|\hat{\boldsymbol{x}}_t^{\mathrm{M1}} - \boldsymbol{x}_t^{\mathrm{M1}}\right\|_2^2 + \left\|\hat{\boldsymbol{x}}_t^{\mathrm{M2}} - \boldsymbol{x}_t^{\mathrm{M2}}\right\|_2^2 + \left\|\hat{\boldsymbol{x}}_t^{\mathrm{F1}} - \boldsymbol{x}_t^{\mathrm{F1}}\right\|_2^2 + \left\|\hat{\boldsymbol{x}}_t^{\mathrm{F2}} - \boldsymbol{x}_t^{\mathrm{F2}}\right\|_2^2\Big) \tag{8}$$

where $\boldsymbol{x}_t^{\mathrm{M1}}$, $\boldsymbol{x}_t^{\mathrm{M2}}$, $\boldsymbol{x}_t^{\mathrm{F1}}$ and $\boldsymbol{x}_t^{\mathrm{F2}}$ are the target LPS features while $\hat{\boldsymbol{x}}_t^{\mathrm{M1}}$, $\hat{\boldsymbol{x}}_t^{\mathrm{M2}}$, $\hat{\boldsymbol{x}}_t^{\mathrm{F1}}$ and $\hat{\boldsymbol{x}}_t^{\mathrm{F2}}$ denote the estimated LPS features for each group at the tth frame. This objective function is optimized via the back-propagation algorithm in a mini-batch mode with T sample frames. Among the references of the four outputs, only two are activated by the underlying two speakers of each input mixture while the other two are set with white Gaussian noises at a 40 dB signal-to-noise-ratio (SNR). Here, due to the logarithm operation to generate LPS features, 40 dB white Gaussian noises were adopted just as a low energy floor to replace the absolute silence to make the training process more stable. In the separation stage, the gender combination of each utterance can be determined using the following rule:

$$\frac{E_{\mathrm{t}}^{\mathrm{M1+M2}}}{E_{\mathrm{t}}^{\mathrm{F1+F2}}} \begin{cases} > \beta & \text{M-M} \\ < 1/\beta & \text{F-F} \\ \in [1/\beta, \beta] & \text{M-F} \end{cases} \tag{9}$$

where $E_{\mathrm{t}}^{\mathrm{M1+M2}}$ and $E_{\mathrm{t}}^{\mathrm{F1+F2}}$ refer to the total energy in the time domain of the utterance level reconstructed from the detector DNN outputs of two male groups and two female groups, respectively. This rule is inspired by the fact that the energy ratio should be extremely high or low for the same-gender mixture and can be confined to a certain SNR range for the different-gender mixture with the SNR threshold β ($\beta > 1$).

C. DNN-Based Speech Separation

In a recent work for speech enhancement [35], DNN was adopted as a regression model to learn the relationship between noisy and clean speech. More recently, a similar architecture was applied to speech separation in supervised or semi-supervised modes [22], [23]. In this paper, we propose a novel gender-combination dependent DNN architecture for unsupervised speech separation as illustrated in Fig. 6. The inputs to DNN are the LPS features of mixed speech with some acoustic context (multiple neighboring frames). The main difference of the proposed framework from the previous efforts [22], [23] is that each output of the DNN is extended to a speaker group rather than a specific speaker. This implies that an unseen speaker could be represented by a speaker group, which is motivated by the analysis in Section II. Furthermore, the use of speaker grouping can alleviate the diversified data collection problem for a target speaker to develop speaker-dependent models [22], [23].

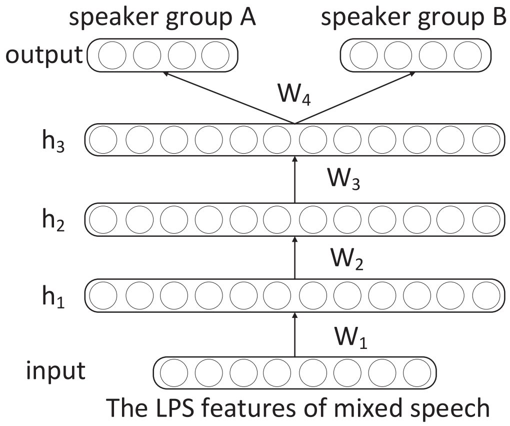

Fig. 6. A DNN-based speech separator, with three of them in total.

Following gender mixture detection, three corresponding DNN separators are designed, one for each decision made in (9). Each DNN separator, illustrated in Fig. 6, adopts dual outputs corresponding to the two speaker groups. The DNN parameters are first initialized randomly and then optimized via minimizing mean squared errors between the DNN outputs and the LPS features of the target and interfering speakers:

$$E_2 = \frac{1}{T}\sum_{t=1}^{T}\left(\parallel \hat{\boldsymbol{x}}_t^{\mathrm{A}} - \boldsymbol{x}_t^{\mathrm{A}} \parallel_2^2 + \parallel \hat{\boldsymbol{x}}_t^{\mathrm{B}} - \boldsymbol{x}_t^{\mathrm{B}} \parallel_2^2\right) \tag{10}$$

where $\hat{\boldsymbol{x}}_t^{\mathrm{A}}$ and $\hat{\boldsymbol{x}}_t^{\mathrm{B}}$ are the estimated LPS features of the two speaker groups, A and B, at the tth frame while $\boldsymbol{x}_t^{\mathrm{A}}$ and $\boldsymbol{x}_t^{\mathrm{B}}$ are the target LPS features. The speaker group combination (A,B) denotes one of three combinations, namely (M,F), (M,M), and (F,F). The implementation detail is similar to (8).

In the separation stage in this study, we only focus on the case that the two mixing speakers are unseen in the training set, which means the segregation is conducted in an unsupervised manner without any information of the underlying speakers. By linking to the discussion in Section II-C, the input mixture could be divided into five categories: M-F, M-M-D, M-M-S, F-F-D, and F-F-S. If the mixing speakers belong to two different groups of (M1, M2, F1, F2), including M-F, M-M-D and F-F-D mixtures, it is reasonable to conduct unsupervised separation using our proposed framework in Fig. 4. However, for the M-M-S and F-F-S mixtures, which are not used for training the detector and separator, the separation performance is unpredictable. This is understandable as separation of the speech mixture with a smaller distance between the two mixing speakers should be more challenging. Our experimental results in Section V provide evidences for these discussions and our approach consistently outperforms the state-of-the-art CASA approach for all input mixture situations.

IV. Experimental Setup

Our experiments were conducted on the SSC corpus [3] with the down-sampled 16 kHz waveforms. The frame length and shift are 512 samples (32 msec) and 256 samples (16 msec), respectively. A short-time Fourier transform [36] was adopted to compute the discrete Fourier transform (DFT) of each overlapping windowed frame. The 257-dimensional LPS features were then used to train DNNs. For waveform reconstruction, the original phase of mixed speech was used with the separated LPS features [37]. Both the detector and separator DNNs consisted of 257*7 $= 1799$ input nodes (a stack of the center frame plus 6 neighbouring frames) and 3 hidden layers with 2048 sigmoidal nodes per layer. The output layer had 514 nodes (dual outputs) for the separator DNNs and 1028 nodes (four outputs) for the detector DNN. The learning rate was set to 0.1 for the first 10 epochs and decreased at a rate of 90% in the next 40 epochs. The mini-batch size T is 128. A global mean and variance normalization scheme to both input and output LPS features was also applied [37]. Moreover, the CDNN consists of 3 hidden layers with 2048 nodes and the output layer with 4 nodes. For the training of CDNN, cross validation set was employed and early-stopping criterion was provided. The separation performance was evaluated using multiple measures, including output SNR [11], sources-to-artifacts ratio (SAR) [38], short-time objective intelligibility (STOI) [39] believed to be highly correlated to speech intelligibility, and perceptual evaluation of speech quality (PESQ) [40] known to have a high correlation with subjective listening scores.

TABLE II
The Speakers Used for Detector and Separator Training

Model			Speaker IDs	
DNN Separators	M-M	M1	1 2 5 10 19 (in blue "○")	
		M2	8 9 14 28 32 (in blue "∗")	
	F-F	F1	18 21 22 23 25 (in red "×")	
		F2	11 16 20 31 33 (in red "□")	
	M-F	M	M1 + M2	
		F	F1 + F2	
DNN Detector		M	M1	M2
		F	F1	F2

A. Training Data Generation

Since there were 16 female and 18 male speakers in total in the SSC corpus, we can only pick small subsets to train the gender mixture detector and speech separators as shown in Figs. 5 and 6. Five speakers were randomly chosen from each of the four groups (M1,M2,F1,F2) as shown in Figs. 2 and 3. The speaker IDs were listed in Table II. The speakers in the M1 and M2 groups were adopted to train M-M separator while the F1 and F2 speakers were used for F-F separator training. The M-F separator was learned with the M1+M2 male group and F1 + F2 female group. All those combinations were finally used to train the gender mixture detector.

Specifically, the input mixture utterances of the training set were generated by randomly adding speaker segments from one group to speaker utterances from another group at SNR levels ranging from -10 dB to 10 dB with an increment of 2 dB. As in this procedure, one speaker utterance should be selected as the fixed reference while other speaker segments are normalized to achieve a specific SNR. The two speakers were exchanged to synthesize the training data in a symmetric manner. For three separators, each DNN was built with 200

TABLE III
THE SPEAKERS USED IN THE TEST SET

Combination		Speaker IDs
M-M	M1	12 27 30 (in blue "○")
	M2	3 6 26 (in blue "∗")
F-F	F1	4 7 15 (in red "×")
	F2	24 29 34 (in red "□")
M-F	M	M1 + M2
	F	F1 + F2

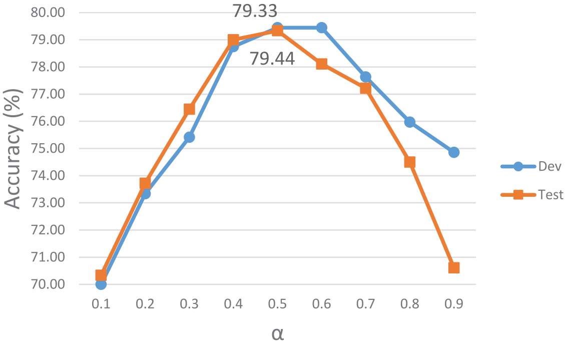

Fig. 7. Performance of GMM-UBM detector.

hours of training data pairs consisting of the input mixture and the underlying speakers. And the gender mixture detector was trained with 100 hours of training data pairs. In this study, the experimental comparisons of different configurations, including the size of training set and SNR coverage, already discussed in [22], [23], were not performed.

B. Testing Data Generation

For the test set, only 3 unseen speakers in the small training set were selected to form each of the four groups, as detailed in Table III. The SNR levels of the test utterances were from -9 dB to 6 dB with an increment of 3 dB. At each SNR level, 300 mixture utterances were synthesized, including 100 M-M, 100 F-F and 100 M-F mixtures. Among the same-gender mixtures (M-M or F-F), one half of the mixing speakers was from different speaker groups (M-M-D or F-F-D) and another half was from the same speaker group (M-M-S or F-F-S).

V. EXPERIMENTS AND RESULT ANALYSIS

A. Experiments on Gender Mixture Detection

To evaluate the performance of the gender mixture detector, a development set similar to the test set, including 40 M-F, 40 M-M and 40 F-F mixtures, was generated for tuning the threshold parameters. The detection accuracies of the development set for the GMM-UBM, CDNN and RDNN detectors were plotted in Figs. 7, 8 and 9, respectively. By searching $\alpha \in [0.1, 0.9]$, $\lambda \in [0.59,0.69]$ and $\beta \in [8, 16]$, optimal values existed for the three techniques to balance the accuracies among different gender combinations. The RDNN detector yielded the best accuracy of 94.86% on the development set, which was significantly better than that of GMM-UBM detector (79.44%) and that of CDNN detector (91.5%). Moreover, a similar observation on the test set could be made from these figures. With the thresholds tuned from the development set, the RDNN detector again achieved the best accuracy of 94.89% while the accuracies of the GMM-UBM and CDNN detector were only 79.33% and 91.11%, respectively.

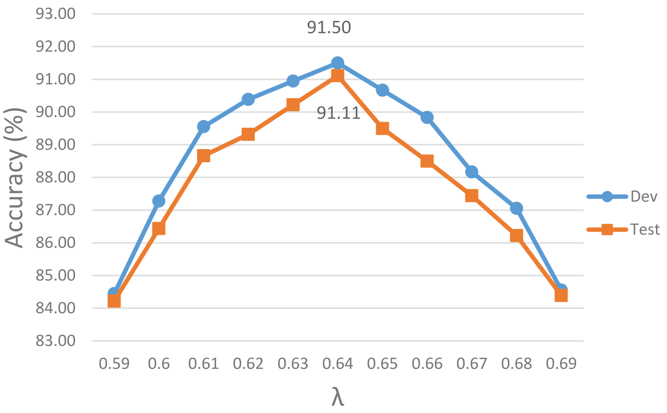

Fig. 8. Performance of CDNN detector.

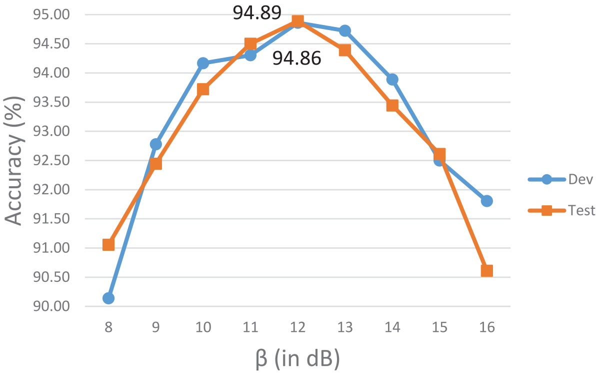

Fig. 9. Performance of RDNN detector.

TABLE IV
THE CONFUSION MATRIX OF THE 600 TESTING UTTERANCES FOR RDNN, CDNN AND GMM-UBM DETECTORS ACROSS ALL SNR LEVELS

Combination	Detector	M-M	F-F	M-F
M-F	RDNN	27	3	570
	CDNN	44	13	543
	GMM-UBM	178	67	355
M-M-D	RDNN	296	0	4
	CDNN	287	0	13
	GMM-UBM	297	3	0
M-M-S	RDNN	291	0	9
	CDNN	280	0	20
	GMM-UBM	298	2	0
F-F-D	RDNN	0	280	20
	CDNN	0	274	26
	GMM-UBM	61	239	0
F-F-S	RDNN	0	271	29
	CDNN	0	256	44
	GMM-UBM	61	239	0

To analyze the detection errors, Table IV presented the confusion matrix of the testing utterances for RDNN, CDNN and GMM-UBM detectors across all SNR levels. Overall, both

TABLE V
OUTPUT SNR (IN DB) COMPARISONS AMONG DIFFERENT DETECTORS

Input SNR (dB)		−9	−6	−3	0	3	6
M-F	Oracle	4.42	5.36	6.32	7.28	8.17	8.97
	RDNN	3.31	5.16	6.32	7.28	8.17	8.84
	CDNN	2.85	4.66	5.83	6.85	7.61	8.22
	GMM-UBM	2.31	3.64	4.56	5.79	5.74	5.64
M-M-D	Oracle	0.41	1.72	3.05	4.20	5.05	5.58
	RDNN	0.41	1.72	3.05	4.15	5.03	5.55
	CDNN	0.37	1.52	3.05	4.11	5.03	5.55
	GMM-UBM	0.41	1.72	3.05	4.20	5.05	5.41
M-M-S	Oracle	−0.27	0.65	1.70	2.75	3.70	4.43
	RDNN	−0.55	0.55	1.65	2.75	3.69	4.42
	CDNN	−0.93	0.25	1.65	2.65	3.69	4.42
	GMM-UBM	−0.27	0.65	1.70	2.72	3.66	4.39
F-F-D	Oracle	1.18	1.46	1.98	2.69	3.29	3.76
	RDNN	0.90	1.25	1.85	2.53	3.26	3.72
	CDNN	0.88	1.25	1.85	2.41	3.02	3.54
	GMM-UBM	0.97	1.27	1.82	2.39	2.90	3.27
F-F-S	Oracle	0.93	1.21	1.88	2.43	2.92	3.39
	RDNN	0.79	1.12	1.80	2.30	2.79	3.07
	CDNN	0.79	1.12	1.76	2.20	2.67	2.99
	GMM-UBM	0.88	1.13	1.69	2.18	2.56	2.95

TABLE VI
STOI AND PESQ COMPARISONS BETWEEN CASA AND THE PROPOSED DNN-BASED APPROACHES FOR M-F MIXTURES OF THE TEST SET

Input SNR (dB)		−9	−6	−3	0	3	6
STOI	CASA	0.58	0.64	0.70	0.74	0.78	0.78
	DNN	0.73	0.83	0.87	0.89	0.91	0.91
PESQ	CASA	0.75	0.93	1.15	1.35	1.54	1.63
	DNN	1.87	2.30	2.55	2.72	2.88	2.98

RDNN and CDNN detectors significantly outperformed the GMM-UBM detector especially for the M-F mixtures due to that the DNN detector was trained in a discriminative manner. For the GMM-UBM detector, the accuracies of the same-gender mixtures were much higher than the different-gender mixtures, which was explained as the likelihood ratio of the same-gender mixture given the two gender GMMs was more distinguishable than that of the different-gender mixture. For RDNN and CDNN detectors, the misclassification cases tended to happen in the F-F and M-F mixtures, particularly for the segments one speaker was masked by another speaker. The detection of the M-M mixtures was the most accurate for all detectors. More interestingly, for misclassification of the same-gender mixtures, the results of both RDNN and CDNN detectors were always poor for the M-F test mixtures with one speaker gender correctly detected while the gender information of two mixing speakers was totally wrong using the GMM-UBM detector. Furthermore, the balances of the errors between the M-F and F-F mixtures for RDNN and CDNN detectors were determined by setting the threshold β and λ, respectively.

Finally, we examined the impact of the gender mixture detector on the separation performance. In Table V the output SNR comparisons among different detectors were given. The "Oracle" system denoted that the correct detection results were provided for the separator selection. One thing to note, in [41] a detector that discriminated all 34 speakers was adopted in the semi-supervised mode when the target speaker was known. This could also be another oracle system as a 100% recognition accuracy on the same SSC corpus of both target and interfering speakers was achieved. However, this study focused on the unsupervised mode. We were not aware of speaker identity and all the speakers in the test set were not included in the training stage. That was why we were not able to conduct the speaker recognition procedure here. Moreover, we only needed to know the speaker group IDs rather than the specific speaker IDs, which was one main motivation of this work.

It was clear that the separation performance gaps between the RDNN detector and the oracle results were quite small in most cases, only with the exception at SNR $= -9$ dB. However, the performance degradation of GMM-UBM detector from the oracle results was significant, especially for the M-F mixtures, e.g., the output SNR decreased from 8.17 dB to 5.47 dB at 3 dB input SNR. The overall results of CDNN were between GMM-UBM and RDNN. CDNN could achieve quite close output SNRs to RDNN for M-M mixtures. However for M-F mixtures, CDNN detector led to a performance degradation especially for high SNR levels, e.g., with output SNR dropping from 8.84 to 8.22 under 6 dB. This implied that the detector played an important role in the subsequent separation and the proposed RDNN detector was quite effective. At -9 dB input SNR, the GMM-UBM detector could achieve an output SNR gain (less than 0.3 dB) over the RDNN detector, e.g. 0.28 dB SNR gain for the M-M-S mixtures, due to the lower RDNN detection accuracy in this case. All those observations could be well explained by the gender mixture detection results shown in Table IV.

B. *Experiments on Speech Separation*

In this subsection, we present the speech separation performance using the proposed detection/separation framework in Fig. 4. To elaborate the separation performance of different input speech mixtures as discussed in Section II-C and III-C, three sets of experiments were designed accordingly. The CASA approach [18] in an unsupervised setting was adopted for performance comparisons.

1) Results on M-F Mixtures: In Table VI, STOI and PESQ comparisons between CASA and the proposed DNN-based approach for the M-F mixture subset of the test set. For these mixtures with different genders, the DNN separator could segregate the speakers with large distances as illustrated in Fig. 1. Although the DNN detector was not perfect, the DNN-based approach consistently outperformed the CASA method for all SNR levels in terms of STOI and PESQ, e.g., STOI from 0.64 to 0.83 and PESQ from 0.93 to 2.30 at -6 dB input SNR.

2) Results on M-M Mixtures: The M-M mixtures, including M-M-D and M-M-S cases, were more challenging to be segregated than the M-F mixtures due to the smaller distance between speakers. Figs. 10 and 11 plot the STOI and PESQ comparisons between CASA and the proposed DNN-based approaches for the M-M mixture subset of the test set. For the DNN-based approach, both the STOI and PESQ of the

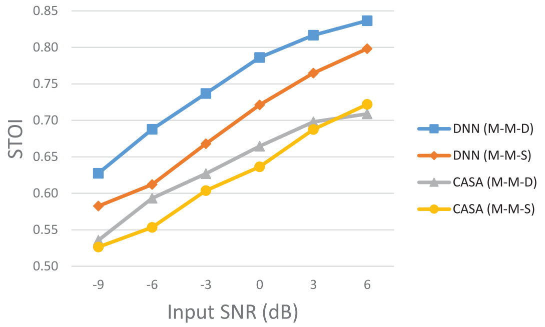

Fig. 10. STOI comparisons between CASA and the proposed DNN-based approaches for M-M mixtures of the test set.

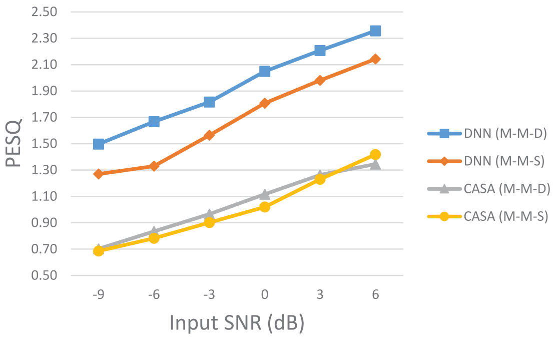

Fig. 11. PESQ comparisons between CASA and the proposed DNN-based approaches for M-M mixtures of the test set.

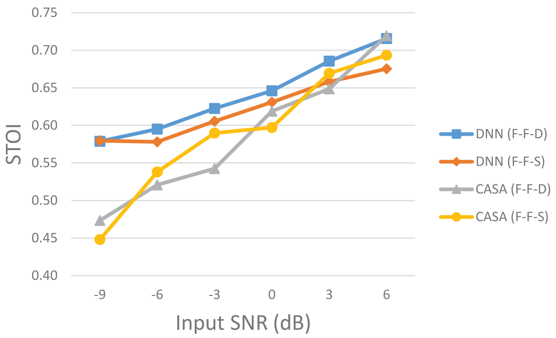

Fig. 12. STOI comparisons between CASA and the proposed DNN-based approaches for F-F mixtures of the test set.

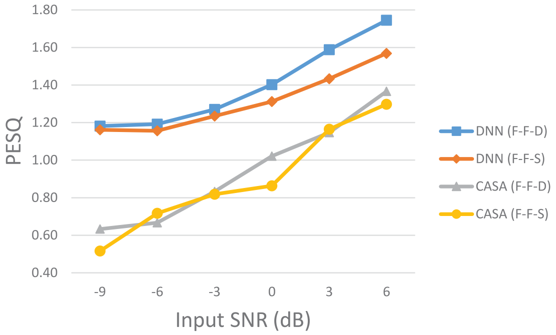

Fig. 13. PESQ comparisons between CASA and the proposed DNN-based approaches for F-F mixtures of the test set.

TABLE VII
STOI COMPARISONS BETWEEN CASA AND DNN-BASED APPROACHES WITH CORRECT DETECTION RESULTS FOR F-F MIXTURES OF THE TEST SET

Input SNR (dB)		−9	−6	−3	0	3	6
F-F-D	CASA	0.47	0.52	0.54	0.62	0.65	0.72
	DNN	0.59	0.61	0.63	0.66	0.69	0.72
F-F-S	CASA	0.45	0.54	0.59	0.60	0.67	0.69
	DNN	0.58	0.58	0.61	0.65	0.67	0.71

M-M-D mixtures were consistently better than those of the M-M-S mixtures and yielded stable performance gains across all SNR levels, e.g., STOI from 0.67 to 0.74 at −3 dB input SNR. In contrast, the CASA approach did not clearly differentiate the M-M-D and M-M-S, giving mixed results, especially for PESQ performance. This implied that the DNN-based approach could be more aware of the speaker dissimilarity by learning from the data. Meanwhile, for both the M-M-D and M-M-S mixtures, the DNN-based approach achieved good performance gains over the CASA method for all SNR levels, e.g., at about a PESQ gain of 1 at 6 dB for the M-M-D mixtures. Even both STOI and PESQ performances for the more difficult M-M-S mixtures using the DNN-based approach could be remarkably better than those of the easier M-M-D mixtures using the CASA technique across all SNR levels.

3) Results on F-F Mixtures: The STOI and PESQ comparisons between CASA and the proposed DNN-based approach for the F-F mixture subset of the test set were displayed in Figs. 12 and 13. By comparing the results of the F-F-D mixtures with F-F-S mixtures, the performance gaps of the DNN-based approach were reduced with the decreased input SNR levels while the results of CASA approach were totally mixed. This could be partially explained as the F-F mixtures were the most challenging cases in comparison to the M-F and M-M mixtures as illustrated in Table I, and the differences between the F-F-D and F-F-S mixtures were not as significant as those between the M-M-D and M-M-S mixtures. In terms of the PESQ performance, the DNN-based approach still consistently outperformed the CASA method for both F-F-D and F-F-S mixtures. However, for the STOI performance, there were two exceptions at 3 dB and 6 dB SNR levels. Based on the experimental analyses, one possible reason was that the detection accuracy of the F-F mixtures was much lower than that of the M-M mixtures as shown in Table IV, leading to the degradation of the STOI performance. If all the detection results were correct, the DNN-based approach could achieve a consistently better (at least the same) STOI performance than the CASA method as shown in Table VII.

Based on the above experiments and analyses, the proposed DNN framework outperformed the CASA approach for most cases of the same-gender mixtures, in terms of both speech quality (PESQ) and intelligibility (STOI) measures. For the DNN-based approach, the M-F mixtures could be well handled while the F-F mixtures were the most challenging cases. This observation was in line with the discussion in Section II-C, i.e., the speaker distances in Table I could well predict the

TABLE VIII
THE OVERALL PERFORMANCE COMPARISONS OF THE CASA, SDNN APPROACH AND THE PROPOSED 2-STAGE DNN-BASED APPROACH AND AVERAGED ACROSS ALL MIXTURE COMBINATIONS OF THE TEST SET AT DIFFERENT SNR LEVELS

Input SNR (dB)		−9	−6	−3	0	3	6
Output SNR	CASA	0.18	1.07	2.19	3.41	4.35	5.34
	2-stage DNN	1.36	2.49	3.50	4.38	5.18	5.74
	SDNN	1.15	1.65	1.73	2.05	2.71	3.54
STOI	CASA	0.52	0.58	0.63	0.67	0.71	0.73
	2-stage DNN	0.64	0.69	0.73	0.76	0.79	0.81
	SDNN	0.62	0.63	0.64	0.64	0.69	0.76
PESQ	CASA	0.67	0.81	0.97	1.12	1.31	1.45
	2-stage DNN	1.47	1.66	1.83	2.00	2.16	2.30
	SDNN	1.29	1.37	1.40	1.48	1.71	2.01
SAR	CASA	−0.92	0.54	1.97	3.42	4.79	5.77
	2-stage DNN	2.18	3.04	3.92	4.82	5.55	6.11
	SDNN	1.75	2.34	2.46	2.59	3.13	3.96

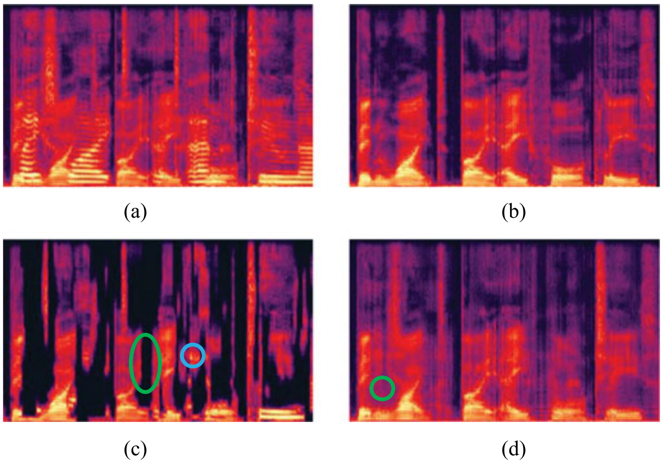

Fig. 14. Spectrograms for an example of the mixture utterance at −6 dB SNR with a male target and a female interferer. (a) Mixed (−6 dB, M + F). (b) Target (M). (c) Unsupervised CASA (M). (d) Unsupervised DNN (M)

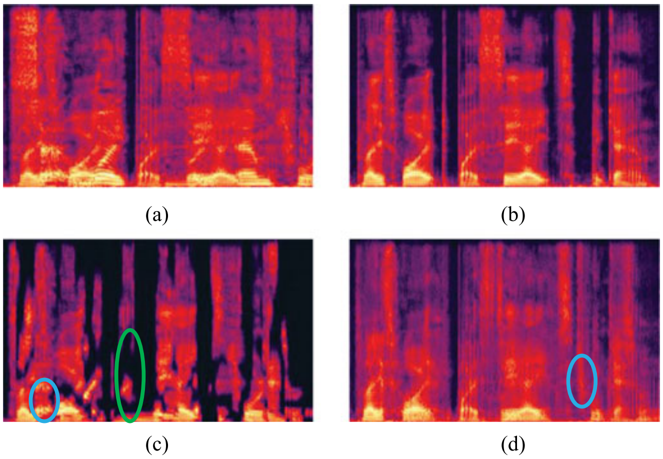

Fig. 15. Spectrograms for an example of the mixture utterance at −3 dB SNR with a male target and a male interferer. (a) Mixed (−3 dB, M + M). (b) Target (M). (c) Unsupervised CASA (M). (d) Unsupervised DNN (M)

separability between speakers. The performance of the DNN-based approach on the challenging F-F mixtures was even comparable to that of the CASA approach on the M-M mixtures. In summary, the DNN-based approach was not only quite effective for M-F/M-M-D/F-F-D mixtures, but also generated good performances on the M-M-S and F-F-S mixtures which were never involved in training the DNN-based detector and separators.

C. Overall Separation Performance Comparisons

Table VIII lists the overall performance comparisons of CASA and the proposed DNN-based approaches averaged across all mixture combinations of the test set at different SNR levels. Four objective measures, including output SNR, STOI, PESQ, and SAR, were adopted in this comprehensive study. Besides our proposed 2-stage DNN-based approach, one single DNN separator (SDNN) with the same architecture in Fig 6 was also designed to accommodate all types of training data, namely M-M, F-F, and M-F mixture utterances. Therefore, the speakers corresponding to the two outputs could be any of the 20 speakers in the training set, which might be a great challenge for the DNN to learn two "confused" speaker groups. So in the separation stage for two unseen mixed speakers, the performance of SDNN seemed unpredictable. Based on the results in Table VIII, several observations could be made. First, 2-stage DNN consistently outperformed SDNN for all objective measures, especially at relatively high SNR levels, e.g., with STOI from 0.69 to 0.79 at 3 dB. This implied that one single DNN indeed could not well accommodate all gender combinations. However, SDNN still significantly improved all measures over CASA at low SNRs, e.g., with PESQ from 0.67 to 1.29 at −9 dB. Second, remarkable improvements were observed by comparing 2-stage DNN-based approach with the CASA approach, especially for STOI and PESQ measures across all input SNR levels, e.g., STOI from 0.67 to 0.76 and PESQ from 1.12 to 2.00 at 0 dB input SNR. This implied that the DNN-based detection/separation framework could improve both the speech quality and intelligibility over the CASA approach. For the output SNR and SAR measures, the gains achieved by 2-stage DNN-based approach over the CASA approach were larger at low SNRs, e.g., output SNR from 1.07 dB to 2.49 dB and SAR from 0.54 dB to 3.04 dB at −6 dB input SNR. These two measures showed that the DNN-based approach simultaneously removed more interferences and generated less artificial signals. All the results verify the effectiveness of the DNN-based detection/separation framework and the reasonable assumption that unseen speakers could be well represented by learning the characteristics of a group of similar speakers. Moreover we believe that in real applications if more diversified speakers and a large mount of training data could be leveraged to learn the detector and separator, more promising performances for unsupervised speech separation would be expected.

Finally, we illustrate some subtle differences of separation results using example spectrograms shown in Figs. 14, 15 and 16. Fig. 14(a) was the spectrogram of a mixture utterance with a male target and female interferer at −6 dB SNR. Fig. 14(b) illustrates the male target. Fig. 14(c) and 14(d) are the spectrograms of separated male target speech using the CASA and DNN-based approaches, respectively. According to [18], the energy normalization was applied to Fig. 14(b), 14(c), and 14(d). Compared with the DNN technique, the CASA approach lost

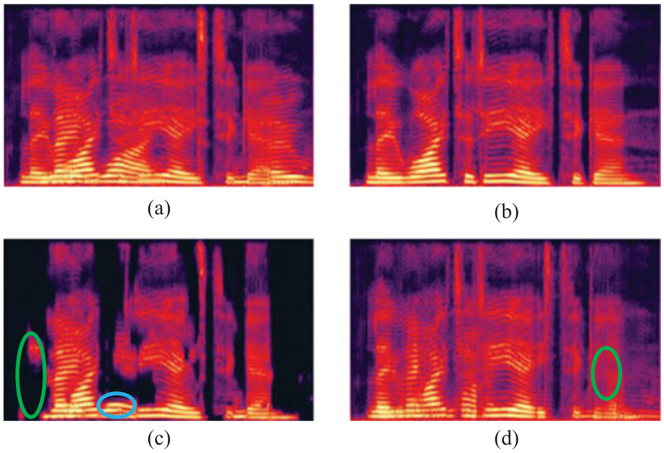

Fig. 16. Spectrograms for an example of the mixture utterance at 0 dB SNR with a female target and a female interferer. (a) Mixed (0 dB, F + F). (b) Target (F). (c) Unsupervised CASA (F). (d) Unsupervised DNN (F)

more target speech details (e.g., in the green circle areas) and preserved more interferences (e.g., in the blue circle areas). It was clear that the separation result of the for the M-F mixture was quite good and close to the reference target speaker. For the same-gender mixtures shown in Figs. 15 and 16, the similar problems with the CASA approach could be observed. Meanwhile, due to the challenge of the same-gender mixtures, especially for the F-F mixtures, the DNN results were also not as good in comparison to the different-gender mixtures, e.g., the remaining interference (blue circle areas) in Fig. 15(d) and lost target information (green circle areas) in Fig. 16(d). In Sections II-C and V-B, a possible reason why the F-F mixtures were more challenging than the M-M mixtures has been discussed. Based on the spectrograms, one factor might be that there were more spectral details lied in the high-frequency bands for the female speech which were more difficult for the DNN detector and separator to learn. Furthermore, the F-F mixture in Fig. 16(a) was also visually more confusable (e.g., with quite similar harmonic structures between speakers) than the M-M mixture in Fig. 15(a). Overall, the DNN-based approach yields more similar spectrograms to the reference target speaker for all mixture combinations. More results and demos can be found at http://home.ustc.edu.cn/˜wyn314/SSC-DNN-USS.html.

VI. Conclusion and Future Work

We propose a novel DNN-based gender mixture detection and speech separation framework for unsupervised single-channel speech separation motivated by the analysis of the speaker dissimilarities. A comprehensive series of experiments and analyses, including the importance of DNN-based detector and the comparisons among different mixture combinations, are conducted. The proposed DNN framework could consistently outperform the state-of-the-art CASA approach in terms of multiple objective measures. This study is a successful demonstration of applying the deep learning technology to unsupervised speech separation in a single-channel setting which is still a challenging open problem. In the future, we aim at refining the proposed framework by designing better speaker grouping algorithms and improving the performance of both detector and separators. Moreover, we plan to further develop our system on larger datasets and even some other languages. The other neural network structures are also going to be explored in the future, such as recurrent neural network for our system. Another interesting direction is to incorporate the dissimilarity measure with cost-functions for DNN-based detector and separator.

References

[1] V. C. Shields, “Separation of added speech signals by digital comb filtering,” M.S. thesis, Dept. Elect. Eng., MIT, Cambridge, MA, USA 1970.

[2] K.-C. Yen and Y. Zhao, “Co-channel speech separation for robust automatic speech recognition: Stability and efficiency,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1997, vol. 2, pp. 859–862.

[3] 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.

[4] S. T. Roweis, “One microphone source separation,” in *Proc. Neural Inf. Process. Syst.*, 2000, vol. 13, pp. 793–799.

[5] S. T. Roweis, “Factorial models and refiltering for speech separation and denoising,” in *Proc. EUROSPEECH*, 2003, pp. 1009–1012.

[6] S. Rennie, J. Hershey, and P. Olsen, “Single-channel multitalker speech recognition,” *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 66–80, Nov. 2010.

[7] 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. 45–66, 2010.

[8] G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Applications to Clustering*. New York, NY, USA: Marcel Dekker, 1988.

[9] 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.

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

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

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

[13] G. J. Mysore, P. Smaragdis, and B. Raj, “Non-negative hidden Markov modeling of audio with application to source separation,” in *Proc. 9th Int. Conf. Latent Variable Anal. Signal Separation*, 2010, pp. 140–148.

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

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

[16] G. Hu and D. 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] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, “A computational auditory scene analysis system for speech segregation and robust speech recognition,” *Comput. Speech Lang.*, vol. 24, no. 1, pp. 77–93, 2010.

[18] K. Hu and D. Wang, “An unsupervised approach to cochannel speech separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 1, pp. 122–131, Jan. 2013.

[19] 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.

[20] O. Dikmen and A. T. Cemgil, “Unsupervised single-channel source separation using Bayesian NMF,” in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2009, pp. 93–96.

[21] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 31–35.

[22] J. Du, Y. Tu, Y. Xu, L. Dai, and C.-H. Lee, “Speech separation of a target speaker based on deep neural networks,” in *Proc. 12th Int. Conf. Signal Process.*, 2014, pp. 473–477.

[23] Y. Tu, J. Du, Y. Xu, L. 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. 9th Int. Symp. Chin. Spoken Lang. Process.*, 2014, pp. 250–254.

[24] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 1562–1566.

[25] X.-L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 5, pp. 967–977, May 2016.

[26] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[27] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proc. INTERSPEECH*, 2009, pp. 1559–1562.

[28] F. Young and R. Hamer, *Theory and Applications of Multidimensional Scaling*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1994.

[29] Y. Zhao, C. Zhang, F. K. Soong, M. Chu, and X. Xiao, "Measuring attribute dissimilarity with HMM KL-divergence for speech synthesis," in *Proc. 6th ISCA Speech Synthesis Workshop*, 2007, vol. 2, pp. 206–210.

[30] J. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Trans. Comput.*, vol. C-18, no. 5, pp. 401–409, May 1969.

[31] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, Oakland, CA, USA., vol. 1, no. 14, 1967, pp. 281–297.

[32] X.-L. Zhang, "Universal background sparse coding and multilayer bootstrap network for speaker clustering," *Proc. INTERSPEECH*, 2016, pp. 1858–1862.

[33] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digit. Signal Process.*, vol. 10, no. 1, pp. 19–41, 2000.

[34] R. B. Dunn, D. A. Reynolds, and T. F. Quatieri, "Approaches to speaker detection and tracking in conversational speech," *Digit. Signal Process.*, vol. 10, no. 1, pp. 93–112, 2000.

[35] 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.

[36] L. Durak and O. Arikan, "Short-time fourier transform: Two fundamental properties and an optimal implementation," *IEEE Trans. Signal Process.*, vol. 51, no. 5, pp. 1231–1242, May 2003.

[37] 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.

[38] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[39] 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. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 4214–4217.

[40] ITU-T "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," International Telecommunication Union-Telecommunication Standardization Sector, Recommendation R. P.862, 2001.

[41] J. Du, Y. Tu, L.-R. Dai, and C.-H Lee, "A regression approach to single-channel speech separation via high-resolution deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 8, pp. 1424–1437, Aug. 2016.

Yannan Wang received the B.S. degree from the the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2011. He is currently working toward the Ph.D. degree at USTC. His research interests include speech separation and language identification.

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, USTC. During this period, he worked as an Intern twice for nine months in Microsoft Research Asia, Beijing, China. In 2007, he also worked as a Research Assistant for six months in the Department of Computer Science, 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, USTC.

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, the M.S. degree from 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, China, in 1997. He joined USTC in 1993. He is currently a Professor in the School of Information Science and Technology, USTC. His 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 (F'97) is a Professor in 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 the Technical Staff, and the Director of the Department of Dialogue Systems Research. He has published more than 450 papers and 30 patents, and was highly cited close to 30 000 times for his original contributions with an h-index of 65 on Google Scholar. Dr. Lee is a Fellow of ISCA. He received numerous awards, including the Bell Labs President's Gold Award in 1998. He also received the IEEE Signal Processing Society'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 International Speech Communication Association Medal in scientific achievement for "Pioneering and seminal contributions to the principles and practice of automatic speech and speaker recognition."