Unsupervised Single-Channel Speech Separation via Deep Neural Network for Different Gender Mixtures

Yannan Wang*, Jun Du*, Li-Rong Dai* and Chin-Hui Lee†
* National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, Anhui, P. R. China
E-mail: wyn314@mail.ustc.edu.cn, jundu@ustc.edu.cn, lrdai@ustc.edu.cn Tel/Fax: 0551-63602575
† School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA. USA
E-mail:chl@ece.gatech.edu

***Abstract*—In this study, we propose a regression approach via deep neural network (DNN) for unsupervised speech separation in a single-channel setting. 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 then proposed to characterize the separation ability between competing speakers. We demonstrate that the distance between speakers of different genders is large enough to warrant a possible separation. We finally propose a DNN architecture with dual outputs, one representing the female speaker group and the other characterizing the male speaker group. Trained and tested on the Speech Separation Challenge corpus our experimental results show that the proposed DNN approach achieves large performance gains over the state-of-the-art unsupervised techniques without using specific knowledge about the mixed target and interfering speakers and even outperforms the supervised GMM-based method.**

I. Introduction

Co-channel speech separation [1], referring to separating a speech component of interest from a noisy mixture, has a variety of important applications, e.g., automatic speech recognition (ASR) [2] in the recent Speech Separation Challenge (SSC) [3]. We can then formulate the problem with two mixing speakers as follows: $\boldsymbol{x}^{\mathrm{m}} = \boldsymbol{x}^{\mathrm{t}} + \boldsymbol{x}^{\mathrm{i}}$, with $\boldsymbol{x}^{\mathrm{m}}$ being the mixed speech signal while $\boldsymbol{x}^{\mathrm{t}}$ and $\boldsymbol{x}^{\mathrm{i}}$ referring to speech of the target and interfering speakers, respectively. Model-based approaches are widely used in speech separation in a supervised mode [4] which generally builds speaker-dependent models assuming the identities of the target and interfering speakers are known. Many approaches to modeling the speakers have been investigated. For instance, Roweis [5] employs the factorial hidden Markov model (FHMM) to learn the information of a speaker and then separate the speech mixture through computing a mask function and refiltering. Another probabilistic model named as factorial-max vector quantization (MAXVQ) is introduced in [6]. The Gaussian mixture model (GMM) is also used in [7], [8] via minimum mean-square estimation (MMSE) to re-synthesize the speech signals. An iterative GMM-based approach is proposed in [9] 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) [10], [11] which decomposes the signal into sets of bases and weight matrices.

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. Therefore at the other extreme, in an unsupervised separation mode, computational auditory scene analysis (CASA) [12], inspired by the ability of human auditory perception to recover signals of interest from background distractions, is widely adopted without assuming any knowledge about mixing speakers. For example, in [13] pitch and amplitude modulation are employed to obtain the voiced components of co-channel speech through grouping estimated pitches. In [14] onset/offset-based segmentation and model-based grouping are introduced to manage unvoiced portions. Unsupervised clustering for sequential grouping is adopted to convert simultaneous streams to two clusters in [15] by maximizing the ratio of between-cluster and within-cluster distances.

In some recent work [16], [17], [18] deep neural network (DNN) is 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. In the supervised speaker-dependent mode, we know both the target and interfering speakers. While in the semi-supervised speaker-independent mode [19], the interferer is assumed unknown. In this study, we extend the DNN approach to unsupervised speech separation of two speakers who are both unknown and relate this feasibility to some speaker distance measures, i.e., the larger the distance between competing speakers the better the mixed speakers could be separated. As one special case, we focus our discussion on segregating speakers with different genders. We propose a deep neural network architecture with dual outputs, one representing the female speaker group and the other characterizing the male speaker group. In other words, our proposed DNN acts like a gender separator to segregate co-channel speech effectively without a need of collecting a set of training utterances for each individual

speaker to be separated. Intuitively, there is a large discrepancy between speakers with different genders, e.g., unique vocal tract, fundamental frequency contour, timing, rhythm, dynamic range, etc. This results in a large distance between male and female speakers in most cases to facilitate a good gender segregation. We evaluate our proposed framework on the SSC corpus [3] and achieve a significantly better separation performance than the state-of-the-art unsupervised techniques and even outperform the GMM-based supervised approach in [9].

II. Characterization of Speaker Distances

In principle some prior information or clues should be leveraged upon for effectively unsupervised speech separation. Here we explore the feasibility of using speaker dissimilarity measures to establish speaker groupings for training DNNs.

In our work we adopt the recently emerged i-vector based speaker representation [20] with the Euclidean distance to measure a speaker dissimilarity as follows:

$$D_2(\mathbf{x}||\mathbf{y}) = \sqrt{\sum_{i=1}^{I}(x_i - y_i)^2} \tag{1}$$

where $\mathbf{x} = (x_1, x_2, ..., x_I)$ and $\mathbf{y} = (y_1, y_2, ..., y_I)$ are I-dimension ($I = 100$) i-vectors for the two speakers trained on the same corresponding SSC data.

To visualize the similarity between two individual objects in a low-dimensional space, each object to be studied can be represented by a point and the points are elaborately arranged in order to approximate the distances between pairs of objects. We adopted multidimensional scaling (MDS) [21] to graphically describe the relationship conveyed by aforementioned distance measurements. The MDS graphs of i-vector based distance matrices for the 34 speakers of the SSC corpus [3] are shown in Fig. 1, respectively. In this figure, the blue and red points represent the male and female speakers. 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. As one demonstration in this study, we focus on the case with different genders, namely one female speaker and one male speaker in mixing. Fig. 1 confirm that the female and the male groups could be well separated in two clusters in most cases, i.e., the distances between speakers of the same gender are smaller than those between speakers of different genders. This motivates our proposed DNN approach in the next section.

III. DNN-based Speech Separation

DNN is essentially a feed-forward multi-layer perceptron with many hidden layers.Recently it has been widely used for classification tasks in image processing and speech recognition. In our recent work for speech enhancement [22], DNN was instead 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

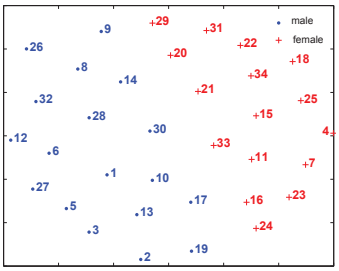

Fig. 1. Multidimensional scaling graph of the i-vector distances among 34 speakers.

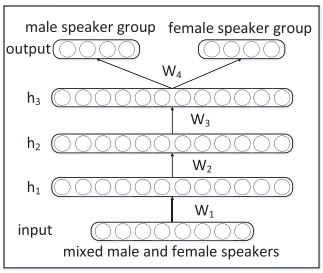

Fig. 2. A DNN architecture for male and female separation.

semi-supervised modes [16], [17]. Motivated by the powerful modeling capability the proposed DNN architecture is illustrated in Fig. 2 with dual outputs for both female and male speaker groups in the current frame given the input features of mixed speech with an acoustic context (multiple neighboring frames). The input of our DNN is mixed by different gender speech of arbitrary speakers while the outputs refer to the separated speech segments of the female and male speaker groups. This architecture avoids the limitations that abundant data of the target speaker is required to develop speaker-dependent models. Besides, in our work the adopted log-power spectral features are capable of providing perceptually relevant parameters. Moreover, our proposed DNN architecture improves the continuity of estimated clean speech along both the time and frequency axes. As a contrast, the conventional GMM-based approach [9] does not well model the temporal dynamics of speech.

Training of DNN is via fine-tuning the network which is initialized randomly achieving a comparable performance with unsupervised pre-training by stacking multiple restricted Boltzmann machinesAnd sigmoid hidden units and linear output units are adopted in our network. Supervised fine-tuning is implemented under the MMSE criterion which jointly minimizes the mean squared error between the DNN outputs and the reference clean features of both the female and male speakers:

$$E = \frac{1}{T}\sum_{t=1}^{T}(\| \hat{\boldsymbol{x}}_t^{\mathrm{m}} - \boldsymbol{x}_t^{\mathrm{m}} \|_2^2 + \| \hat{\boldsymbol{x}}_t^{\mathrm{f}} - \boldsymbol{x}_t^{\mathrm{f}} \|_2^2) \tag{2}$$

where $\| \cdot \|_2$ refers to L2 norm and T is the total number of training frames in a mini-batch. $\hat{\boldsymbol{x}}_t^{\mathrm{m}}$ and $\hat{\boldsymbol{x}}_t^{\mathrm{f}}$ are the estimated features of the male and female targets at frame t. Moreover,

$\boldsymbol{x}_t^{\mathrm{m}}$ and $\boldsymbol{x}_t^{\mathrm{f}}$ are the reference clean features of the male and female targets, respectively.

In the training stage, all the information about the speakers selected for the female and male groups are known as a supervised mode. However, in the separation stage, both the female and male speakers are unseen which corresponds to the unsupervised speech separation.

IV. Experiments and Results

For evaluation we randomly selected 100 different gender mixtures consisting of 50 male-female and 50 female-male combinations from the whole SSC test set, referred to as two-talker mixtures with signal-to-noise-ratios (SNRs, here we consider interfering speech as noise) ranging from -9dB to 6dB with an increment of 3dB. We built 5 DNNs in total which were trained on different speaker groups as shown in Table I. All the utterances of 10 male and 10 female speakers in the training set were used to train each DNN. It was then evaluated on the speech mixtures of the other unseen 8 male and 6 female speakers. And the input utterances were created by randomly

TABLE I
SPEAKERS USED IN THE TRAINING STAGE.

model	speaker IDs	
DNN1	male	1 2 3 5 6 8 9 10 12 13
	female	4 7 11 15 16 18 20 21 22 23
DNN2	male	1 2 3 8 12 14 17 19 26 28
	female	4 7 11 15 21 24 25 29 31 33
DNN3	male	1 2 9 10 13 14 27 28 30 32
	female	4 16 18 20 22 24 25 29 31 33
DNN4	male	1 5 6 10 17 19 26 27 30 32
	female	7 11 15 16 18 21 23 24 33 34
DNN5	male	3 5 6 9 12 13 17 19 26 27
	female	4 11 16 20 21 22 23 24 25 29

adding segments of one speaker to another speaker with a different gender at SNRs ranging from -10dB to 10dB with an increment of 2dB. This gave a good coverage of SNRs in the test set. Moreover, the training mixture utterances were generated in an asymmetric manner, namely fixing one speaker as the target and normalizing the energy of another speaker as the interferer to achieve a specific SNR. To balance the two genders, one half of the training data was generated with male speakers as targets and the other half with female target speakers.

We down-sampled the original waveforms from 25kHz to 16kHz. The frame length and shift are 512 samples (32 msec) and 256 samples (16 msec), respectively. A short-time Fourier transform was adopted to compute the discrete Fourier transform (DFT) of each overlapping windowed frame. Then 257-dimensional log-power spectral features were used to train DNNs. For the waveform reconstruction, the original phase of mixed speech was adopted with the separated log-power spectra [23]. In all experiments the DNN consisted of 1799 input nodes (stack of 7 neighboring frames), 3 hidden layers which used a sigmoid activation function with 2048 nodes per layer, and 514 output nodes (estimated features of male and

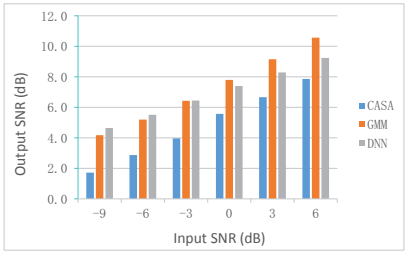

Fig. 3. Output SNR comparison of different gender separation approaches.

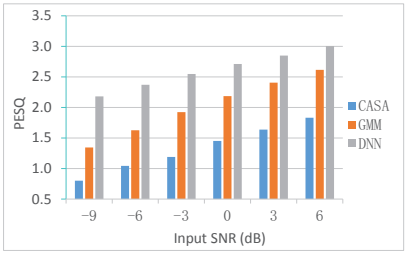

Fig. 4. PESQ comparison of different gender separation approaches.

female targets). The global normalization was also applied to the input features to have zero mean and unit variance.

A. Result Comparisons of Separation Experiments

Next, iterative GMM [9] in a supervised scenario (denoted as "GMM") and CASA using segmentation and grouping [15] in an unsupervised setting (denoted as "CASA") were adopted for performance comparisons. In Figures 3 and 4 we plot bar charts of the two competing techniques in terms of output SNR and PESQ [24], respectively. Apparently, the supervised GMM-based approach yielded a significant improvement over unsupervised CASA for both output SNR and PESQ. Similar improvements were also observed comparing the proposed DNN with the unsupervised CASA approach, e.g., PESQ rising from 1.04 to 2.37 at an input SNR of -6 dB. Moreover, our proposed unsupervised DNN framework also outperformed the GMM-based supervised approach in PESQ across all input SNR levels, e.g., PESQ increasing from 1.63 to 2.38 at an input SNR of -6 dB. Furthermore, we also found that our proposed DNN obtained a SNR gain over GMM under -9dB and -6dB SNR conditions, while the performance was worse than GMM when SNR is greater than 0dB.

By considering that our DNN-based approach was operating in the unsupervised mode while the GMM-based system was running in the supervised mode, those results were very encouraging for designing practical algorithms of enhancing the speech quality in co-channel speech separation. The above improvements also demonstrated that using multiple speakers could well simulate the characteristics of unseen female and male mixing speakers. Other speaker grouping could be further explored based on some newly defined or existing speaker dissimilarity measures.

Finally, we illustrate some detailed separation performance with an utterance example in Fig. 5. Fig. 5(a) is the spectro-

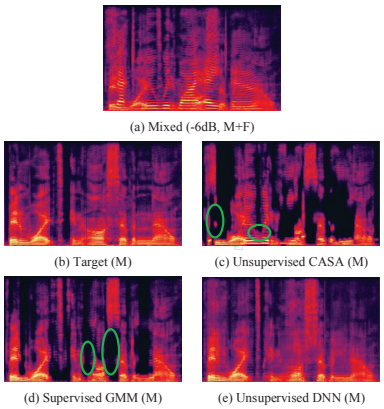

Fig. 5. Illustration of spectrograms for separating the target male utterance from the mixed utterance with a female interferer at -6 SNR.

gram of the mixture with a male target and female interferer at -6dB and Fig. 5(b) refers to the male target. Fig. 5(c) is the spectrogram of separated male target speech with the unsupervised method based on CASA while Fig. 5(d) is the result of the GMM-based method. Fig. 5(e) is the spectrogram of the separated male target with our proposed DNN approach. All the results are normalized to promote the energy level. It is observed that the CASA approach lost many target speech details and still preserved some interference speech at low and high frequency as shown in the green circles while this problem was alleviated with GMM-based approach as speaker dependent models were constructed. But the GMM-based approach was still disturbed by speech details lost as shown in the green circle of Fig. 5(d). In contrast, our proposed DNN approach could achieve the most similar spectrograms to the reference of target speaker. More results and demos can be found at http://home.ustc.edu.cn/w̃yn314/SSC_DNN.html.

V. Conclusion and Future Work

In this study we demonstrate that in unsupervised co-channel speech separation the proposed DNN framework could well segregate speech from mixtures of two unseen speakers with different genders. Significant performance improvements have been achieved when compared with unsupervised techniques based on CASA. Furthermore our proposed approach also outperforms GMM-based separation operating in a supervised mode in terms of speech quality although the output SNR gains are lower under relatively high SNR conditions. Although the different gender mixtures are relatively easier to be handled, our DNN architecture design is a quite reasonable demonstration with the evidence of speaker dissimilarity measures. As for future work, we will investigate how to extend our approach to the same gender cases.

Acknowledgment

This work was partially funded by the National Nature Science Foundation of China (Grant No. 61273264 and No. 61305002), the Electronic Information Industry Development Fund of China (Grant No. 2013-472), and the Programs for Science and Technology Development of Anhui Province, China (Grants No. 13Z02008-4 and No. 13Z02008-5).

References

[1] V. C. Shields, "Separation of added speech signals by digital comb filtering," *S.M. Thesis, Dept. of Electrical Engineering, MIT*, 1970.
[2] K.-C. Yen and Y. Zhao, "Co-channel speech separation for robust automatic speech recognition: stability and efficiency," in *Proc. ICASSP*, vol. 2, 1997, 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] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "Speaker-independent model-based single channel speech separation," *Neurocomputing*, vol. 72, no. 1, pp. 71–78, 2008.
[5] S. T. Roweis, "One microphone source separation," in *NIPS*, vol. 13, 2000, pp. 793–799.
[6] ——, "Factorial models and refiltering for speech separation and denoising," in *Proc. Eurospeech*, 2003, pp. 1009–1012.
[7] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1766–1776, 2007.
[8] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2299–2310, 2007.
[9] K. Hu and D. Wang, "An iterative model-based approach to cochannel speech separation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 14, 2013.
[10] M. Schmidt and R. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. INTERSPEECH*, 2006, pp. 2614–2617.
[11] J. Le Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in *Proc. ICASSP*, 2015.
[12] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.
[13] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.
[14] 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.
[15] K. Hu and D. Wang, "An unsupervised approach to cochannel speech separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 122–131, 2013.
[16] 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. ICSP*, 2014, pp. 473–477.
[17] 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. ISCSLP*, 2014, pp. 250–254.
[18] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. ICASSP*, 2014, pp. 1562–1566.
[19] M. Zhrer, R. Peharz, and F. Pernkopf, "Representation learning for single-channel source separation and bandwidth extension," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2398–2409, Dec 2015.
[20] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
[21] F. Young and R. Hamer, "Theory and applications of multidimensional scaling," *Hillsdale, NJ: Eribaum Associates*, 1994.
[22] Y. Xu, J. Du, L. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.
[23] ——, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions. Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan 2015.
[24] ITU-T and R. P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs ," *International Telecommunication Union-Telecommunication Standardisation Sector*, 2001.