A CASA APPROACH TO DEEP LEARNING BASED SPEAKER-INDEPENDENT CO-CHANNEL SPEECH SEPARATION

Yuzhou Liu[1] *and DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA
{liuyuz,dwang}@cse.ohio-state.edu

ABSTRACT

We address speaker-independent co-channel speech separation from the computational auditory scene analysis (CASA) perspective. Specifically, we decompose the two-speaker separation task into the stages of simultaneous grouping and sequential grouping. Simultaneous grouping is first performed at the frame level by separating the spectra of two speakers with a permutation-invariantly trained recurrent neural network (RNN). In the second stage, the simultaneously separated spectra at each frame are sequentially grouped into the utterances of the two underlying speakers by a clustering RNN. Overall optimization is then performed to fine tune the two-stage system. The proposed CASA approach takes advantage of permutation invariant training (PIT) and deep clustering (DC), but overcomes their shortcomings. Experiments show that the proposed system improves over the best reported results of PIT and DC.

Index Terms— Co-channel speech separation, computational auditory scene analysis, deep learning, permutation invariant training, deep clustering

1. INTRODUCTION

Co-channel speech separation refers to the task of separating speech of two simultaneous speakers in a monaural recording. As an important branch of the cocktail party problem, co-channel speech separation is useful for a wide variety of speech applications, e. g., meeting transcription, speaker identification, and hearing aids. Although human auditory systems are extremely good at focusing at one speaker in the presence of interfering speakers, this problem has stayed largely unsolved for machines for more than 5 decades.

Before the deep learning era, a common approach to co-channel speech separation is computational auditory scene analysis (CASA) [19]. In CASA, auditory cues, like onset and pitch, are utilized to group sound sources in both frequency and time domain, known as simultaneous grouping and sequential grouping, respectively. For example, a tandem algorithm [10] groups time-frequency (T-F) representation of speech into coherent T-F segments by using iterative pitch estimation and mask estimation. Hu and Wang [11] improve the tandem algorithm by further introducing clustering based sequential grouping for T-F segments. Besides CASA, model based approaches, e. g., non-negative matrix factorization (NMF) [17] and GMM-HMM [16], have also been explored. However, they only introduce moderate performance gain due to limited modeling capabilities.

This research was supported in part by an NIDCD grant (R01 DC012048) and the Ohio Supercomputer Center.

With the rapid development of deep learning, more and more researchers start to formulate co-channel speech separation as a regression problem. The general idea is to feed spectral features into deep neural networks (DNNs) to predict T-F masks or spectra of two speakers in a mixture [3, 12, 20]. There are usually two output layers in DNNs, one for the target speaker and the other for the interfering speaker. Most studies [3, 12, 20] use a matched target speaker during training and test, denoted by target-dependent DNNs. It has been shown that this technique leads to substantial intelligibility improvement for hearing impaired listeners [7]. However, if such a DNN is trained on an open set of speakers with random layouts of labels, severe permutation problem may happen. For example, utterance 1 and 2 are both mixtures of speaker A and B. If the labels in utterance 1 are organized as (A, B), and the labels in utterance 2 are organized as (B,A), then conflicting gradients may be generated for the two mixtures, and prevent the DNN from converging.

The frame-level permutation invariant training (denoted by tPIT) algorithm [15] solves this problem by looking at all possible label permutations within each frame during training, and only uses the one with the lowest frame-level loss to update the network. A locally optimized output-target pairing can thus be reached, which leads to very promising frame-level separation performance. However, the speaker assignment in tPIT's output may swap frequently across frames. To further address this issue, an utterance-level PIT (uPIT) algorithm [15] is proposed, which uses bi-directional long short-term memory (BLSTM) recurrent neural networks (RNNs) to perform sequence-to-sequence mapping. It forcedly aligns each speaker to a fixed output layer throughout a whole training utterance.

In the meanwhile, deep clustering (DC) [8] tackles the permutation problem by training a BLSTM-RNN to assign an embedding vector to each T-F unit of the spectrogram, such that embedding vectors of T-F units dominated by the same source are similar, and embedding vectors of those dominated by different sources have larger distances. Clustering these embedding vectors using the K-means algorithm assigns each T-F unit to one of the speakers in the mixture, which can be viewed as a binary mask for speech separation. In [13], several upgrades, including deeper network, recurrent dropout and end-to-end training are proposed to improve DC. In [2], a concept of attractors is further introduced to DC to enable end-to-end training.

Although uPIT and DC work well in speaker-independent situations, they both have drawbacks. As reported in [15], uPIT sacrifices frame-level performance to achieve better assignment at the utterance level. Its sequence-to-sequence mapping mechanism works poorly for same-gender speaker pairs. On the other hand, DC estimates an embedding vector for every T-F unit in an utterance, which is an overkill and is inefficient during inference. Moreover, the im-

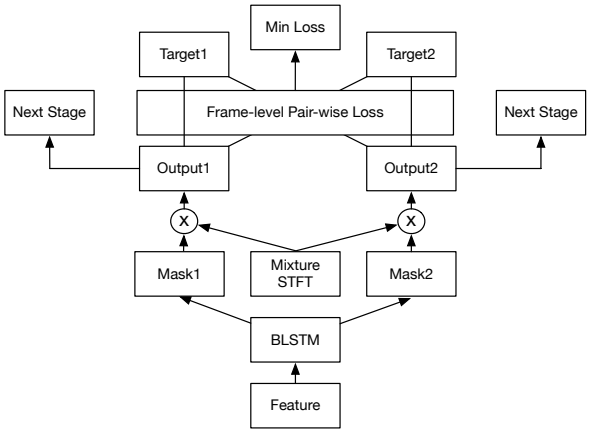

Fig. 1: Diagram of the simultaneous grouping stage.

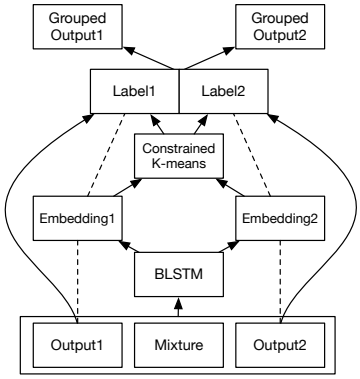

Fig. 2: Diagram of the sequential grouping stage.

plication of an embedding vector is ambiguous when the two underlying speakers have similar energies.

Inspired by PIT, DC and CASA, we propose a deep learning based CASA approach to perform speaker-independent co-channel speech separation. The CASA approach consists of two stages, a simultaneous grouping stage and a sequential grouping stage. In the first stage, a tPIT-BLSTM-RNN is trained to predict the spectra of the two speakers at each frame, with unknown speaker assignment. This stage separates different frequency components of the two speakers at a frame-level, which corresponds to simultaneous grouping in CASA. In the sequential grouping stage, a concatenation of the two estimated spectra and the mixture spectrum is fed into another BLSTM-RNN to predict embedding vectors for the estimated spectra, such that the embedding vectors corresponding to the same speaker are close together, and those corresponding to different speakers are far apart. A constrained K-means algorithm is then performed to forcedly assign the two spectrum predictions at the same frame to different speakers. This stage corresponds to sequential grouping in CASA, which streams short speech segments according to temporal continuities and similarities. The overall system takes advantage of accurate frame-level estimation of tPIT and overcomes the T-F ambiguity issue of DC by applying constrained K-means clustering on frame-level spectra. In the end, we explore end-to-end optimization of the two stages to further fine-tune the system.

In the remainder of this paper, we present the simultaneous grouping stage in Section 2. The sequential grouping stage is described in Section 3. In Section 4, we present experimental results and comparisons. A conclusion is given in Section 5.

2. SIMULTANEOUS GROUPING STAGE

Co-channel speech separation aims to separate two concurrent speakers in a single-microphone recording. Inspired by auditory scene analysis (ASA) [1] and CASA [19], we decompose this task into two stages, a frequency-domain simultaneous grouping stage, and a time-domain sequential grouping stage. This section explains the first stage in details.

Let $X_i(t,f)$ denote the short-time discrete Fourier transform (STFT) of speaker i (i =1, 2), where t and f are the time and frequency indices. The mixture of the two speakers can be defined as:

$$Y(t,f) = X_1(t,f) + X_2(t,f) \tag{1}$$

Conventional deep learning based co-channel speech separation systems [12, 15] feed the magnitude STFT of the mixture signal $|Y(t,f)|$ into a neural network to predict a T-F mask $M_i(t,f)$ for each speaker i. The masks are then multiplied with the mixture signal to reconstruct the original sources:

$$|\tilde{X}_i(t,f)| = M_i(t,f) \odot |Y(t,f)|, \quad i = 1,2 \tag{2}$$

Here $\odot$ denotes element-wise multiplication, and $|\tilde{X}_i(t,f)|$ is the reconstructed magnitude STFT of the i^{th} speaker. $|\tilde{X}_i(t,f)|$ is then coupled with noisy phase to resynthesize the time-domain signal of each speaker.

Various training targets of $|\tilde{X}_i(t,f)|$ have been explored for masking based speech separation in [4]. The phase-sensitive approximation (PSA) is found to be the best one since it can make up errors introduced by the noisy phase during resynthesis. In PSA, the desired reconstructed signal, i.e., the training target, is defined as: $|X_i(t,f)| \odot cos(\phi_i(t,f))$, where $\phi_i(t,f)$ is the element-wise phase difference between the mixture $Y(t,f)$ and the source $X_i(t,f)$. Overall, the training loss at each frame is computed as:

$$J_t = \sum_{f=1}^{F} \sum_{i=1}^{2} ||M_i(t,f) \odot |Y(t,f)| - |X_i(t,f)| \odot cos(\phi_i(t,f))||_2^2 \tag{3}$$

where $||\cdot||_2$ denotes l_2 norm.

Using a predefined ordering of the two targets may cause severe permutation problem for different speaker pairs [15], and may stop the network from converging. Frame-level PIT (tPIT) is proposed to overcome this issue, where targets are provided as a set instead of an ordered list, and the output-target pairing $i \leftrightarrow \theta_i(t)$, for a given frame t, is defined as the pairing that minimizes the loss function over all possible speaker permutations P. In tPIT, the training loss at each frame can be rewritten as:

$$J_t^{tPIT} = \min_{\theta_i(t) \in P} \sum_{f,i} ||M_i \odot |Y| - |X_{\theta_i(t)}| \odot cos(\phi_{\theta_i(t)})||_2^2 \tag{4}$$

We omit (t,f) in M, Y, X, and ϕ notations for simplicity.

tPIT does a great job separating the two speakers at a frame level [15]. Therefore, we directly adopt tPIT as our simultaneous grouping module. A diagram of our tPIT module is illustrated in Fig. 1. In the network, BLSTM is used to exploit temporal context. The targets and outputs are paired w.r.t. the minimum frame-level loss. Finally, the two reconstructed outputs are fed to the next stage for

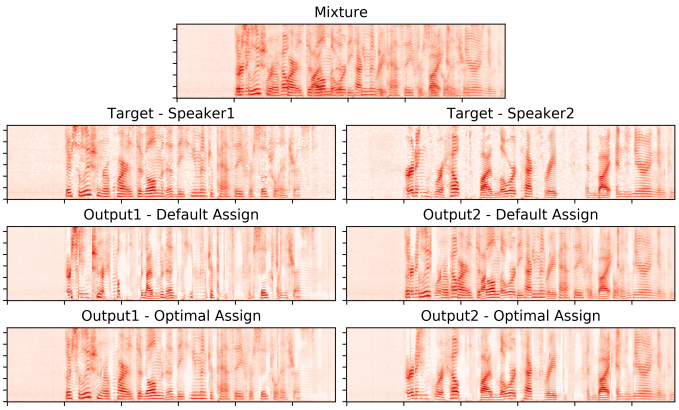

Fig. 3: Results of the simultaneous grouping stage for a male-male test mixture. First row: magnitude STFT of the mixture. Second row: magnitude STFT of the two targets. Third row: two outputs with the default speaker assignment. Last row: two outputs with the optimal speaker assignment.

sequential grouping. Fig. 3 shows an application of tPIT for a male-male test mixture. Due to tPIT's locally optimized training criterion, the output-to-speaker assignment changes very often in tPIT's default output (third row), and it is nowhere close to the targets on the utterance level. However, if we reassign the outputs w.r.t. the minimum loss for each speaker, tPIT can almost perfectly reconstruct both signals, as shown in the last row.

Such optimal speaker assignments can only be achieved when the training targets are known beforehand, which is irrational for real applications. To address this issue, the utterance-level PIT (uPIT) is proposed to perform separation and speaker tracing simultaneously. In uPIT, the output-target pairing $i \leftrightarrow \theta_i(t)$ is fixed for a whole utterance, which corresponds to the pairing that provides the minimum utterance-level loss over all possible permutations. As reported in [15], uPIT greatly improves the separation quality without knowing the optimal speaker assignment. However, the fixed output pairing for an whole utterance prevents the frame-level loss to be optimized as in tPIT. The large gap between same-gender and different-gender performance reported in uPIT [15] also implies that a better tracing algorithm can be designed.

In the proposed CASA approach, we make use of the frame-level outputs of tPIT and propose a new sequential grouping method to track them through time. The details are explained in the next section.

3. SEQUENTIAL GROUPING STAGE

3.1. Deep Clustering Network

In this stage, we trace all frame-level reconstructed spectra using a deep clustering network, which corresponds to sequential grouping in CASA. Deep clustering based speech separation is first proposed by Hershey et al. in [8]. We make a few modifications to DC to make it work under the new setup.

A diagram of our sequential grouping network is illustrated in Fig. 2. Outputs from tPIT $|\tilde{X}_i(t,f)|$ are concatenated with $|Y(t,f)|$ to form the input to the network. Batch normalization is applied afterwards. The network uses BLSTM to project each frame-level output from tPIT $|\tilde{\mathbf{X}}_i(t)|$, into a D-dimensional embedding space $\mathbf{V}_i(t) \in \mathbb{R}^D$, where $|\tilde{\mathbf{X}}_i(t)|$ denotes the vector representation of the

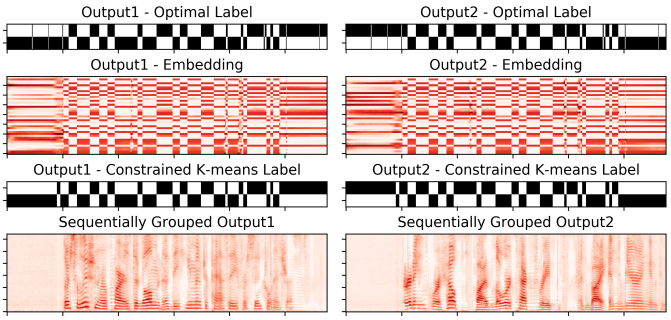

Fig. 4: Results of the sequential grouping stage for the same male-male test mixture as in Fig. 3. All horizontal axes correspond to time. First row: optimal speaker assignment label of the two output layers. Second row: estimated embedding vectors of the two output layers. Third row: estimated speaker assignment label after constrained K-means clustering. Last row: sequentially grouped outputs using K-means labels.

i^{th} output at frame t, and $\mathbf{V}_i(t)$ is the corresponding unit-length embedding vector. The target or optimal label of the network is a two-dimensional indicator vector, denoted by $\mathbf{A}_i(t)$. During the training of tPIT, if the minimum loss is achieve when $|\tilde{\mathbf{X}}_i(t)|$ is paired with speaker 1, we set $\mathbf{A}_i(t)$ to [1 0], otherwise $\mathbf{A}_i(t)$ is set to [0 1]. In other words, $\mathbf{A}_i(t)$ indicates the optimal speaker assignment of $|\tilde{\mathbf{X}}_i(t)|$. $\mathbf{V}_i(t)$ and $\mathbf{A}_i(t)$ can be reshaped into a $2T \times D$ matrix $\mathbf{V}$ and a $2T \times 2$ matrix $\mathbf{A}$, respectively, to represent embedding and assignment information of all frames in an utterance. A permutation independent loss function between $\mathbf{V}$ and $\mathbf{A}$ is presented as:

$$J^{DC} = ||\mathbf{V}\mathbf{V}^T - \mathbf{A}\mathbf{A}^T||_F^2 \tag{5}$$

where $||\cdot||_F$ is Frobenius norm. Optimizing J^{DC} forces $\mathbf{V}_i(t)$ corresponding to the same speaker to get closer during training, and $\mathbf{V}_i(t)$ corresponding to different speakers to become further apart. Since we only care about the speaker assignment of spectra with significant energies, a binary weight for each frame-level output is used during training, only retaining those embedding vectors whose corresponding outputs have greater energy than some ratio (set to -30 dB) of the maximum frame-level energy.

In the next step, a constrained K-means algorithm is used to cluster $\mathbf{V}_i(t)$ into two groups. First, an initial centroid pair is selected as two embedding vectors at a same frame, but with the largest distance in between. Three iterations of K-means are then performed for embedding vectors with significant energies. In the end, we assign the two embedding vectors at each frame to different clusters, making sure that the minimum distance is achieved at the frame level.

After the K-means algorithm, we look up the two outputs from tPIT, and stream frame-level outputs with the same K-means label into one speaker. Results of this stage are shown in Fig. 4, where the same male-male speaker pair is used as in Fig. 3. In Fig. 4, the estimated embedding vectors and the resulting K-means labels almost perfectly match the patterns in the optimal speaker assignment labels. Consequently, the sequentially grouped outputs also match the optimally assigned outputs in Fig. 3.

The major difference between DC and the proposed CASA approach is that, DC's embedding-clustering framework is performed on a T-F level, whereas in CASA it is performed on a Time-Speaker level. There are several advantages to do it in our way. First, estimating an embedding vector for a T-F unit with similar energies from the two speakers is usually error-prone. We avoid this problem by clus-

tering frame-level spectra that can well separate the two speakers. Second, during the clustering stage, we reduce the computational complexity of DC from $O(FT)$ to $O(2T)$. Last, our framework is more flexible. Sequential grouping can be utilized to group various output targets, including different mask types, or even time-domain waveforms. However, DC can only be applied on the T-F domain.

3.2. End-to-end Training

End-to-end (E2E) training can be applied to the CASA approach by using uPIT based training loss on the final sequentially grouped outputs. To do so, we fix all tunable parameters in the embedding-clustering network, and only use them to generated K-means labels for uPIT training. As the training goes, the simultaneous grouping module will be tuned to have higher synergy with the sequential grouping module, and better utterance-level performance can be achieved.

Moving forward, we can train the two stages of the CASA approach in an iterative fashion. However, due to the time limitation, we leave this as a future work.

4. EVALUATION AND COMPARISON

4.1. Experimental Setup

We conduct experiments on the co-channel speech separation corpus introduced in [8], which has a 30-hour training set and a 10-hour validation set generated by selecting random speaker pairs in the Wall Street Journal (WSJ0) training set si_tr_s, and mixing them at various signal-to-noise ratios (SNRs) between 0 dB and 5 dB. Evaluation is conducted on the 5-hour test set, which is similarly generated using 16 unseen speakers from the WSJ0 development set si_dt_05 and si_et_05. All mixtures are sampled at 8 kHz. Magnitude STFT is used as the input feature in both stages, with a frame length of 32ms, a frame shift of 8 ms, and the square root of hanning window is applied. We report our results in terms of signal-to-distortion ratio improvement (SDRi) [18], a metric widely used in speech enhancement evaluation.

4.2. Models

The tPIT network in the simultaneous grouping stage contains 3 BLSTM layers, with 896×2 units in each layer. Two output layers with the ReLU activation function [6] are then used to predict phase sensitive masks. The network is trained with the Adam optimization algorithm [14] and dropout regularization [9]. To accelerate training, we do not apply recurrent dropout [5] to our model, despite its effectiveness for small to median training sets [5, 13]. The initial learning rate is set to 0.0002, and we decrease the learning rate by a ratio of 0.8 when the cross-validation loss stops decreasing for over 8 epochs.

The sequential grouping module has a 4-layer BLSTM-RNN with 300×2 units in each layer. The embedding dimension is set to 40, thus we use 80 sigmoid output units to predict the two embedding vectors at each frame. The initial learning rate is set to 0.001 in this module. Other training recipes follow those in the simultaneous grouping stage, exactly. In both stages, BLSTM-RNNs are trained on the whole utterance level. tPIT-BLSTM is trained first. We then fix parameters in tPIT-BLSTM and start to update the sequential grouping network. In the end, end-to-end training is performed to further fine tune the algorithm.

In addition, we trained a uPIT model with the same configuration as the tPIT for comparison.

Table 1: SDRi (dB) comparison of tPIT and uPIT in terms of default and optimal speaker assignment.

	Optimal Assign	Default Assign
tPIT	12.3	-1.6
uPIT	11.4	10.3

Table 2: SDRi (dB) comparison of different systems with different gender combinations.

	Same Gender	Different Gender	Overall
uPIT [15]	7.5	12.2	10.0
DAN [2]	-	-	10.5
DC++ [13]	9.4	12.0	10.8
CASA	9.5	12.2	10.9
CASA-E2E	9.6	12.2	11.0
tPIT-OPT	11.9	12.6	12.3

4.3. Results and Comparisons

In Table 1, we compare a uPIT-BLSTM with the tPIT module in our system, with exactly the same training recipes. Although the uPIT model improves the results of the default assignment by a large margin, there is still a significant gap between uPIT and tPIT when the optimal output assignment is used. Therefore, we use tPIT as the basis, and come up with a better tracking algorithm to improve PIT.

In Table 2, we compare the proposed CASA approach with other state of the art algorithms, including uPIT [15], deep attractor network (DAN) [2] and the updated version of deep clustering (DC++) [13], in terms of same-gender, different-gender and overall SDRi. The best published result is given for each comparison method. As shown in the table, uPIT has the best performance for the different-gender pairs, but its same-gender results are way worse than other systems, indicating that sequence mapping does not work well for challenging speaker pairs. The proposed CASA approach outperforms uPIT and DAN, and yields slightly better results than DC++. The updated version of the CASA approach, CASA-E2E, is trained by conducting 10 epochs of end-to-end training, which further improves the SDRi of CASA. The upper bound of the proposed approach is the optimally assigned tPIT, as reported in the last row of Table 2. Although CASA-E2E has close SDRi to tPIT-OPT on different-gender pairs, it still can not match the same-gender results. More improvement can be expected when iterative training, recurrent dropout, and smoother tracking algorithm are added to the system in the future.

5. CONCLUSION

We have proposed a CASA approach for deep learning based speaker-independent co-channel speech separation. A simultaneous grouping stage is first conducted to separate the two speakers at the frame-level. Sequential grouping is then performed to stream frame-level spectra into two sources based on their similarities and continuities. The two stages can be trained jointly and iteratively. The proposed CASA approach takes advantage of both permutation-invariant training and deep clustering, and has been shown to yield better results than both approaches. Future work includes incorporating recurrent dropout, iterative joint optimization and real-time processing.

6. REFERENCES

[1] A. Bregman, *Auditory scene analysis*. Cambridge MA: MIT Press, 1990.

[2] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. ICASSP*, 2017, pp. 246–250.

[3] J. Du, Y. Tu, Y. Xu, L. R. Dai, and C. H. Lee, "Speech separation of a target speaker based on deep neural networks," in *Proc. ICSP*, 2014, pp. 65–68.

[4] H. Erdogan, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*, 2015, pp. 708–712.

[5] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Adv. NIPS*, 2016, pp. 1019–1027.

[6] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *AISTATS*, 2011, pp. 315–323.

[7] E. W. Healy, M. Delfarah, J. L. Vasko, B. L. Carter, and D. L. Wang, "An algorithm to increase intelligibility for hearing impaired listeners in the presence of a competing talker," *J. Acoust. Soc. Amer.*, vol. 141, pp. 4230–4239, 2017.

[8] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[9] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[10] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, pp. 2067–2079, 2010.

[11] K. Hu and D. L. Wang, "An unsupervised approach to cochannel speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 122–131, 2013.

[12] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. ICASSP*, 2014, pp. 1562–1566.

[13] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. INTERSPEECH*, 2016, pp. 545–549.

[14] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[15] M. Kolbæk, D. Yu, Z. H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 1901–1913, 2017.

[16] T. T. Kristjansson, J. R. Hershey, P. A. Olsen, S. J. Rennie, and R. A. Gopinath, "Super-human multi-talker speech recognition: the IBM 2006 speech separation challenge system," in *Proc. INTERSPEECH*, 2006, pp. 97–100.

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

[18] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 1462–1469, 2006.

[19] D. L. Wang and G. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley-IEEE Press, 2006.

[20] X. L. Zhang and D. L. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, pp. 967–977, 2016.