Utterance-level Permutation Invariant Training with Discriminative Learning for Single Channel Speech Separation

Cunhang Fan[1,3]*, Bin Liu*[1]*, Jianhua Tao*[1,2,3]*, Zhengqi Wen*[1]*, Jiangyan Yi*[1]*, Ye Bai*[1,3]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
{cunhang.fan, liubin, jhtao, zqwen, jiangyan.yi, baiye}nlpr.ia.ac.cn

Abstract

The challenge in deep learning for speaker independent speech separation comes from the label ambiguity or permutation problem. Utterance-level permutation invariant training (uPIT) technique solves this problem by minimizing the mean square error (MSE) over all permutations between outputs and targets. It is a state-of-the-art deep learning architecture. However, uPIT only minimizes the chosen permutation with the lowest MSE, not discriminates it with other permutations. This may lead to increase the possibility of remixing the separated sources. In this paper, we propose a uPIT with discriminative learning (uPIT-DL) method to solve this problem by adding one regularization at the cost function. In other words, we minimize the difference between the outputs of model and their corresponding reference signals. Moreover, the dissimilarity between the prediction and the targets of other sources is maximized. We evaluate the proposed model on WSJ0-2mix dataset. Experimental results show 22.0% and 24.8% relative improvements under both closed and open conditions compared with the uPIT baseline.

Index Terms: Utterance-level Permutation Invariant Training, Discriminative Learning, BLSTM, Single Channel Speech Separation

1. Introduction

The human auditory system can easily segregate their interested speech in a complex acoustic environment [1, 2, 3]. However, such a problem, known as the cocktail party problem [4], seems to be extremely difficult for machine.

Over the decades, many attempts have been made to solve the cocktail-party problem. Inspired by the human auditory, computational auditory scene analysis (CASA) [5] is proposed to explain perceptual grouping of regions in terms of their similarity [6]. Non-negative matrix factorization (NMF) [7] is another popular technique, which aims to learn a set of nonnegative bases that can be used to estimate mixing factors during evaluation. However, both CASA and NMF not only rely on accurate trackers but also increase computational complexity, they have led to very limited success in multi-talker speech separation [8].

In recent years, deep learning has emerged as a powerful learning method which achieves state-of-the-art performance in many applications. Motivated by the success of deep learning, many deep learning based methods are proposed for speech separation recently [9, 10, 11, 12, 13]. However, for speaker-independent multi-talker speech separation, the difficulty comes from the label ambiguity or permutation problem. To solve this problem, deep clustering (DPCL) [14] is proposed and achieves competitive results. In DPCL, a bidirectional long-short term memory (BLSTM) network is trained to assign contrastive embedding vectors to each time-frequency (TF) bin of the spectrogram. During evaluation, TF bins of different speakers are clustered by using K-means to obtain speaker-dependent mask. And then masks are applied on the mixed signals to acquire the speech of individual speakers. However, researchers assume that each TF bin only belongs to one speaker in DPCL, which is inappropriate. Another shortcoming of DPCL is that they define the objective function in the embedding space, which is sub-optimal because it can't train the model end-to-end. To overcome this limitation, the deep attractor network (DANet) [15] method is proposed. Its network forms attractor points in a high-dimensional embedding space of the signal, and the similarity between attractors and T-F embeddings is converted into a soft separation mask. Unfortunately, the limitation of deep attractor is that it has to estimate attractor points during test.

Recently, to solve the speaker-independent speech separation problem with end-to-end training [16], permutation invariant training (PIT) [17] and utterance level PIT (uPIT) [18] are proposed. The PIT method solves the label ambiguity problem by minimizing the permutation with the lowest MSE at frame level, but it does not solve the speaker tracing problem. To solve the problem, the uPIT is proposed. With uPIT, however, the permutation corresponding to the minimum utterance-level separation error is used for all frames in the utterance. In other words, the pair-wise scores are computed for the whole utterance assuming all output frames follow the same permutation. Therefore, uPIT doesn't need to additional speaker tracing step during inference. However, uPIT only minimizes the chosen permutation with the lowest MSE, does not discriminates it with the other permutations. This may lead to increase the possibility of remixing the separated sources.

The high signal-to-interference ratio (SIR) can improve the intelligibility of speech separation. Therefore, to have high SIR and separate the speech better, many researchers explore discriminative objective function by adding one regularization at the cost function [19, 20, 21]. Motivated by those, in this paper, we propose a uPIT-DL for single channel speaker-independent speech separation. We add one regularization at the cost function following the uPIT, which is called discriminative learning. We minimize the difference between the outputs of model and their corresponding reference signals. Furthermore, we maximize the dissimilarity between the chosen permutation and the other permutations. In other words, it discriminates the target speaker with the other speakers. This helps in decreasing the

possibility of remixing the separated sources and also achieves better separation for the estimated sources.

We evaluate the proposed method uPIT-DL on WSJ0-2mix dataset. Experimental results show that uPIT-DL compares favorably to conventional uPIT and generalizes well over unseen speakers. In other words, the proposed method uPIT-DL is a speaker independent speech separation architecture, which is similar to human.

The rest of this paper is organized as follows. In section 2, masking based monaural speech separation is presented. The details of the proposed model are described in section 3. Section 4 introduces the experimental setup and results. Finally, conclusions are drawn in Section 5.

2. Masking based monaural speech separation

Monaural speech separation aims at separating a linearly mixed single channel microphone signal $y(t)$ into individual source signals $x_s(t)$, $s = 1, ..., S$, where S is the number of source signals. The relationship of the mixed signal $y(t)$ and source signals $x_s(t)$ can be represented as:

$$y(t) = \sum_{s=1}^{S} x_s(t) \tag{1}$$

The corresponding short-time Fourier transformation (STFT) of those signals are $Y(t, f)$ and $X_s(t, f)$. The following relationship is still satisfied after STFT

$$Y(t, f) = \sum_{s=1}^{S} X_s(t, f) \tag{2}$$

Our aim is to estimate each source signal $x_s(t)$ from $y(t)$ or $Y(t, f)$. It is well-known that mask based speech separation can obtain a better result [22]. According to the commonly used masking method, the estimated magnitude $|\widetilde{X}_s(t, f)|$ of each source can be estimated by

$$|\widetilde{X}_s(t, f)| = |Y(t, f)| \odot M_s(t, f) \tag{3}$$

where $\odot$ indicates element-wise multiplication and $M_s(t, f)$ is the mask of source s. It is very difficult to estimate phase directly for speech separation and speech enhancement. Therefore, the estimated magnitude $|\widetilde{X}_s(t, f)|$ and the phase of mixed signal are used to reconstruct each source signal by inverse STFT (ISTFT).

3. Discriminative learning system based on uPIT

In this paper, we propose a uPIT-DL system to estimate the mask of each source as shown in Figure 1. First, the uPIT model chooses the permutation with the lowest MSE. Then, a regularization is added at the cost function following the uPIT, which is called discriminative learning. With the discriminative learning, the differences between the chosen permutation and the other permutations are maximized. This helps in decreasing the possibility of remixing the separated sources and achieves better separation for the estimated sources.

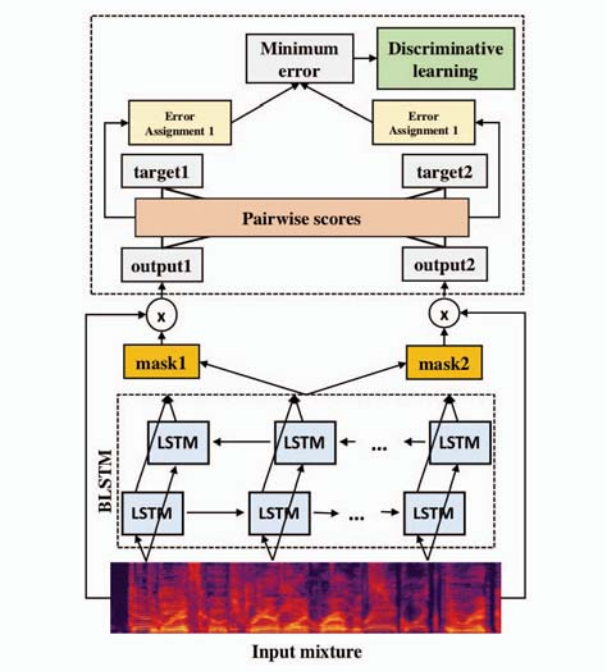

Figure 1: *The proposed uPIT-DL system architecture.*

3.1. Mask and uPIT

We use the BLSTM network to estimate the mask of each source signal. The ideal amplitude mask (IAM) which is a widely-accepted mask for speech separation [18, 23] is utilized in this paper. And it is defined as

$$M_s^{IAM}(t, f) = \frac{|X_s(t, f)|}{|Y(t, f)|} \tag{4}$$

The IAM can achieve the highest SDR when the phase of each source equals the phase of the mixed speech. However, in most cases this assumption is unreasonable. Although IAMs satisfy the constraint that $0 \leq M_s^{IAM}(t, f) < \infty$, the majority of the TF units are in the range of $0 \leq M_s^{IAM}(t, f) \leq 1$ [18]. Therefore, the output activation functions for estimating IAMs can be softmax, sigmoid and ReLU.

As for uPIT, the training criterion is the MSE between the estimated magnitude and true magnitude.

$$J_{iam} = \frac{1}{TF} \sum_{s=1}^{S} |||Y| \odot \widetilde{M}_s - |X_s|||_F^2 \tag{5}$$

where TF is the total number of TF units over all sources and $||.||_F$ is the Frobenius norm.

The uPIT solves the label ambiguity problem by choosing the minimal cost among all permutations (P).

$$\widetilde{p} = \underset{p \in P}{arg\,min}\, J_{mask} \tag{6}$$

where J_{mask} is the J_{iam}.

3.2. Discriminative learning

For uPIT, minimizing Eq.6 is to make the predictions and the targets more similar. However, the high SIR can improve the

intelligibility of speech separation. Therefore, we explore discriminative objective function that not only increase the similarity between the prediction and its target, but also decrease the similarity between the prediction and the targets of other sources.

The discriminative learning maximizes the dissimilarity between the chosen permutation and the other permutations by adding a regularization at the cost function. The cost function of the proposed model is defined as

$$J = \widetilde{p} - \sum_{p_i \neq \widetilde{p}, p_i \in P} \lambda_i p_i \tag{7}$$

where p_i is a permutation from P but not $\widetilde{p}$, $\lambda_i \geq 0$ is the regularization parameter of p_i. When $\lambda_i = 0$, the proposed uPIT-DL is same as the uPIT.

For two-talker speech separation, we assume that p_1 is the permutation with the lowest MSE. Therefore, when the IAM is used, the cost function becomes as follow:

$$\begin{aligned} J &= p_1 - \lambda p_2 \\ &= \frac{1}{TF}\sum(||Y \odot \widetilde{M}_1 - X_1||_F^2 - \lambda||Y \odot \widetilde{M}_1 - X_2||_F^2 \\ &+ ||Y \odot \widetilde{M}_2 - X_2||_F^2 - \lambda||Y \odot \widetilde{M}_2 - X_1||_F^2) \end{aligned} \tag{8}$$

From Eq.8 we can know that the discriminative learning enlarges the distance of the target source with the other sources. It means that it maximizes the differences between the target speakers with the others.

Therefore, the proposed model with discriminative learning minimizes the difference between the outputs of model and their corresponding reference signals. Simultaneously, it maximizes the dissimilarity between the target source and the others. So the discriminative learning decreases the possibility of remixing the separated sources and also achieves better separation.

4. Experiments and Results

4.1. Datasets

We evaluate the methods on the WSJ0-2mix dataset [14], which is derived from WSJ corpus [24]. From the WSJ0 training set si_tr_s, the 30h training set (20,000 utterances) and the 10h validation set (5,000 utterances) consisting of two-speaker mixtures are generated by randomly selecting utterances by different speakers (50 male and 51 female speakers), and mixing them at various signal-to-noise ratios (SNR) between 0dB and 5dB. From the WSJ0 development set si_dt_05 and evaluation set si_et_05, the five hours test set (3,000 utterances) is generated similarly using utterances from 10 male and 8 female speakers.

Speakers in the validation set are the same as those in the training set, so we use the validation set to select the best model from all epochs and to evaluate the source separation performance in closed conditions (CC). Moreover, because the speakers in the test set are different from those in the training set and validation set, the test set is considered as open condition (OC) evaluation.

4.2. Experimental setup

The sampling rate of all generated data is 8 kHz before processing to reduce computational and memory costs. The 129-dim log spectral magnitudes of the mixed speech are used as the input features, which are computed using a short-time Fourier transform (STFT) with 32 ms length hamming window and 16 ms window shift. The magnitudes of two targets are generated in the same way.

In this paper, to train the model more quickly and reduce the memory costs, 3 BLSTM layers with 128 units in each layer are deployed. However, there is 896 units in [18]. The mask estimation layer uses sigmoid as the activation function. All models contain random dropouts with a dropout rate 0.5. Each minibatch contains 20 randomly selected utterances. The number of epoch is set to 50. Our models are optimized with Adam algorithm [25] and implemented using Tensorflow deep learning framework [26]. The learning rate is adjusted adaptively by Adam algorithm.

4.3. Evaluation metric

In this work, in order to quantitatively evaluation speech separation results, the models are evaluated on signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) which are the BBS-eval [27] score. And they are widely used to evaluate speech separation performance. The SDR is defined as follow:

$$SDR = 10log_{10}\frac{||s_{target}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2} \tag{9}$$

where $s_{target} = f(s_j)$ is a version of the source s_j modified by an allowed distortion $f \in \mathcal{F}$, and where e_{interf}, e_{noise} and e_{artif} are respectively the interferences, noise and artifacts error terms.

The SIR is defined as follow:

$$SIR = 10log_{10}\frac{||s_{target}||^2}{||e_{interf}||^2} \tag{10}$$

The normalized SDR (NSDR) is defined as:

$$NSDR(\widetilde{x}, x, y) = SDR(\widetilde{x}, x) - SDR(y, x) \tag{11}$$

where $\widetilde{x}$ is the separated speech, x is the original clean speech, and y is the mixture.

The same as the NSDR, the normalized SIR (NSIR) is defined as:

$$NSIR(\widetilde{x}, x, y) = SIR(\widetilde{x}, x) - SIR(y, x) \tag{12}$$

We report the overall performance via global NSDR (GNSR) and global NSIR (GNSIR) which are the weighted means of the NSDRs and NSIRs (NSIR) respectively, over all test clips weighted by their length. The higher values of SDR and SIR represent the better separation performance.

4.4. Experimental results

4.4.1. uPIT-DL vs. conventional uPIT

Table 1 shows the result of GNSDR (same as "SDR improvement" in [17, 18]) and GNSIR (dB) improvement comparisons between our proposed method and the uPIT approach on the WSJ0-2mix database.

From table 1 we can make several observations. First, when λ is 0.1, as for GNSDR, our proposed method uPIT-DL achieves a better performance than conventional uPIT($\lambda = 0$). To be specific, the GNSDR of our proposed method uPIT-DL ($\lambda = 0.1$) is 6.6 dB on closed condition and 6.7 dB on open condition. However, as for the conventional uPIT($\lambda = 0$), the

GNSDR is 6.4 dB on both closed and open condition. This indicates that our proposed method uPIT-DL can obtains a better performance for speech separation. Since the discriminative learning separates the target speaker with others, it provides a constraint to ensure that the output frames of the same speaker do not remix to the other speakers.

Second, from the GNSIR scores we can know that the discriminative learning can significantly improve the performance of speech separation compared with conventional uPIT. Both the λ 0.1 and 0.3, the proposed method all perform better than uPIT. Especially, when $\lambda = 0.3$, the proposed method achieves 12.2 dB and 12.5 dB GNSIR in closed and open condition. But as for uPIT, it is only 10.0 dB and 10.1 dB. Compared with uPIT, a further 22.0% and 24.8% relative improvement is obtained under both closed and open condition. It shows the effectiveness of the proposed method.

Table 1: *The result of GNSDR and GNSIR (dB) for different separation methods on the WSJ0-2mix dataset with optimal assignment on closed (CC) and open (OC) condition. λ is the regularization parameter of the proposed method*

Method	Mask	λ	GNSDR		GNSIR	
			CC	OC	CC	OC
uPIT	IAM	0	6.4	6.4	10.0	10.1
uPIT-DL	IAM	0.1	**6.6**	**6.7**	11.4	11.5
uPIT-DL	IAM	0.3	6.1	6.3	**12.2**	**12.5**

As an example, Figure 2 shows the spectrogram for a single two-speaker test case along with spectrograms of clean speech signal, uPIT model as well as the proposed model with different λ. Notice that, compared with the spectrograms of clean speech signals, the harmonics of the proposed model are preserved well, and the formant structures are seen to be effectively preserved in the reconstructed speech. Those indicate that the mixed signal is effectively separated by the proposed model. On the other hand, uPIT can separate different speech stream from mixed speech, but the formant structure is not clear compared with the proposed method. For example, compared with (c)(d) in Figure 2, we can see that in (b) some formant structures are not reconstructed and some mixed speech is not separated very well (marked in the black boxes).

4.4.2. uPIT-DL with different λ

From table 1 we can know that when λ gets larger, the value of GSIR increases. This indicates that the proposed method can improve the value of SIR for speech separation. However, when λ is 0.3, the uPIT-DL achieves worse performance than uPIT for GNSDR. The reason is that after several epochs, the value of $\widetilde{p}$ in Eq.9 is very small and closes to zero so that the cost function becomes as follow:

$$J \approx -\sum_{p_i \neq \widetilde{p}, p_i \in P} \lambda_i p_i \tag{13}$$

There is no information about the target sources. If the value of λ is too large, the model would maximize the dissimilarity between the target source and the others only. Therefore, it will lead to over-fitting.

5. Conclusions

In this paper, we propose a discriminative objective function in uPIT for single channel speaker independent multi-talker speech separation. Different from the conventional uPIT model, our proposed model not only minimizes the chosen permutation with the lowest MSE, but also discriminates it with the other permutations. In other words, we minimize the difference between the outputs of model and their corresponding reference signals. Simultaneously, we maximize the dissimilarity between the prediction and the targets of other sources. The experimental results show that our proposed model achieves 22.0% and 24.8% relative improvements under both closed and open conditions comparing with the uPIT baseline. This demonstrates that the discriminative learning can improve the performance of speech separation. In the future, we will explore the combination of pitch with the proposed method for multi-channel speech separation.

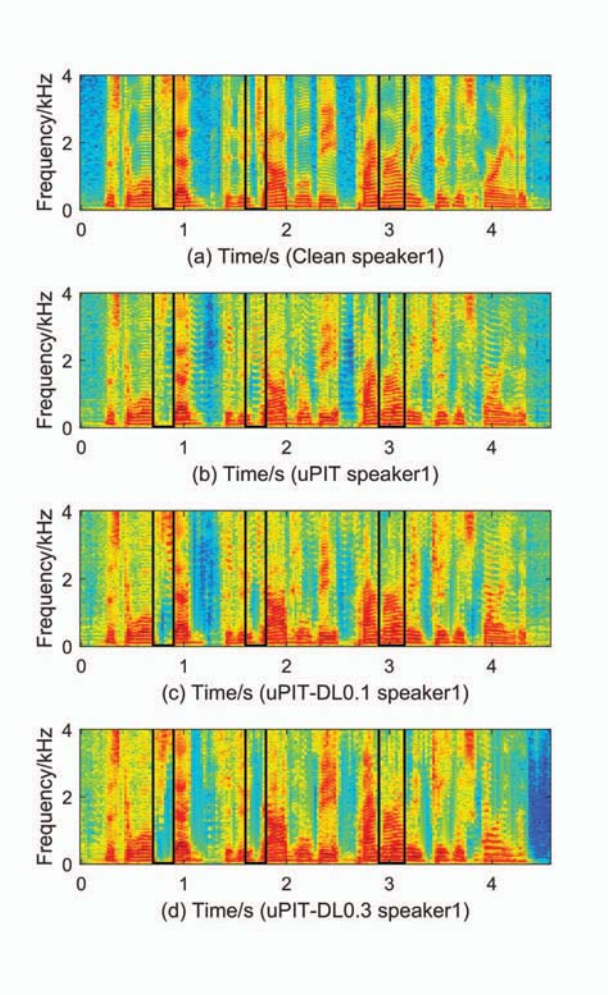

Figure 2: *An example of target, conventional uPIT, the proposed system with λ 0.1 and 0.3 spectrogram for a speech segment from the test set. (a):spectrogram of the target speech signal; (b):uPIT($\lambda = 0$); (c):uPIT-DL0.1 ($\lambda = 0.1$); (d):uPIT-DL0.3 ($\lambda = 0.3$).*

6. Acknowledgements

This work is supported by the National Key R&D Program of China (No.2017YFB1002802) and the National Natural Science Foundation of China (NSFC) (No.61425017, No.61773379, No.61332017, No.61603390, No.61771472).

7. References

[1] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinncunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial eeg," *Cerebral Cortex*, vol. 25, no. 7, p. 1697, 2015.

[2] M. Cooke, *Modelling auditory processing and organisation*. Cambridge University Press, 1993.

[3] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 1996.

[4] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution." *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[5] C. Darwin, "Computational auditory scene analysis: Principles, algorithms and applications," *IEEE Transactions on Neural Networks*, vol. 19, no. 1, pp. 199–199, 2008.

[6] M. Wertheimer, "Laws of organization in perceptual forms." *Psychologische Forschung*, vol. 4, pp. 71–88, 1950.

[7] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *INTERSPEECH 2006 - Icslp, Ninth International Conference on Spoken Language Processing, Pittsburgh, Pa, Usa, September*, 2006.

[8] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech & Language*, vol. 24, no. 1, pp. 1–15, 2010.

[9] J. Zegers *et al.*, "Multi-scenario deep learning for multi-speaker source separation," in *Proceedings ICASSP 2018*, 2018, pp. 5379–5383.

[10] L. Li and H. Kameoka, "Deep clustering with gated convolutional networks," in *Proceedings ICASSP 2018*, 2018, pp. 16–20.

[11] C. Li, L. Zhu, S. Xu, P. Gao, and B. Xu, "Cbldnn-based speaker-independent speech separation via generative adversarial training," in *Proceedings ICASSP 2018*, 2018, pp. 711–715.

[12] D. Wang and J. Chen, "Supervised speech separation based on deep learning: an overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[13] Y. Wang, J. Du, L. R. Dai, and C. H. Lee, "A maximum likelihood approach to deep neural network based nonlinear spectral mapping for single-channel speech separation," in *INTERSPEECH*, 2017, pp. 1178–1182.

[14] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 31–35.

[15] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," *international conference on acoustics, speech, and signal processing*, pp. 246–250, 2017.

[16] L. H. Xu C, Xiao X *et al.*, "Single channel speech separation with constrained utterance level permutation invariant training using grid lstm," in *Proceedings ICASSP 2018*, 2018, pp. 6–10.

[17] D. Yu, M. Kolbæk, Z. H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 241–245.

[18] M. Kolbæk, D. Yu, Z.-H. Tan, J. Jensen, M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1901–1913, 2017.

[19] E. M. Grais, G. Roma, A. J. R. Simpson, and M. D. Plumbley, "Combining mask estimates for single channel audio source separation using deep neural networks," in *Proc INTERSPEECH 2016, 8-12 Sep*, 2016, pp. 3339–3343.

[20] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks." in *ISMIR*, 2014, pp. 477–482.

[21] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 1562–1566.

[22] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 22, no. 12, pp. 1849–1858, 2014.

[23] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 708–712.

[24] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

[26] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, and M. Devin, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," 2016.

[27] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.