INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Discriminative Learning for Monaural Speech Separation Using Deep Embedding Features

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

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

Abstract

Deep clustering (DC) and utterance-level permutation invariant training (uPIT) have been demonstrated promising for speaker-independent speech separation. DC is usually formulated as two-step processes: embedding learning and embedding clustering, which results in complex separation pipelines and a huge obstacle in directly optimizing the actual separation objectives. As for uPIT, it only minimizes the chosen permutation with the lowest mean square error, doesn't discriminate it with other permutations. In this paper, we propose a discriminative learning method for speaker-independent speech separation using deep embedding features. Firstly, a DC network is trained to extract deep embedding features, which contain each source's information and have an advantage in discriminating each target speakers. Then these features are used as the input for uPIT to directly separate the different sources. Finally, uPIT and DC are jointly trained, which directly optimizes the actual separation objectives. Moreover, in order to maximize the distance of each permutation, the discriminative learning is applied to fine tuning the whole model. Our experiments are conducted on WSJ0-2mix dataset. Experimental results show that the proposed models achieve better performances than DC and uPIT for speaker-independent speech separation.

Index Terms: deep clustering, uPIT, speech separation, discriminative learning, deep embedding features

1. Introduction

Monaural speech separation aims to estimate target sources from mixed signals in a single-channel. It is a very challenging task, which is known as the cocktail party problem [1].

In order to solve the cocktail party problem, many works have been done over the decades. Traditional speech separation methods include computational auditory scene analysis (CASA) [2], non-negative matrix factorization (NMF) [3] and minimum mean square error (MMSE) [4]. However, these methods have led to very limited success in speaker-independent speech separation [5].

Recently, supervised methods using deep neural networks have significantly improved the performance of speech separation [6, 7, 8, 9, 10, 11]. Deep clustering (DC) [12] is a deep learning based method for speech separation and achieves impressive results. It trains a bidirectional long-short term memory (BLSTM) network to map the mixed spectrogram into an embedding space. At testing stage, the embedding vector of each time-frequency (TF) bin is clustered by K-means to obtain binary masks. However, the objective function of DC is defined in the embedding space, which can't be trained end-to-end. To overcome this limitation, the deep attractor network (DANet) [13] method is proposed. DANet creates attractor points in a high-dimensional embedding space of the acoustic signals. Then the similarities between the embedded points and each attractor are used to directly estimate a soft separation mask at the training stage. Unfortunately, it enables end-to-end training while still requiring K-means at the testing stage. In other words, it applies hard masks at testing stage.

The permutation invariant training (PIT) [14] and utterance-level PIT (uPIT) [15] are proposed to solve the label ambiguity or permutation problem of speech separation. The PIT method solves this problem by minimizing the permutation with the lowest mean square error (MSE) at frame level. However, it does not solve the speaker tracing problem. To solve this problem, uPIT is proposed. With uPIT, the permutation corresponding to the minimum utterance-level separation error is used for all frames in the utterance. Therefore, uPIT doesn't need speaker tracing step during inference. However, uPIT and PIT only use the mixed amplitude spectrum as input features, which can't discriminate each speaker very well. In addition, uPIT doesn't increase the distance between the chosen permutation and others. This may lead to increasing the possibility of remixing the separated sources. In [16], a Chimera network [17] is applied to speech separation, which uses a multi-task learning architecture to combine the DC and uPIT. However, it simply employs the DC and uPIT as two outputs rather than fusion with each other.

In this paper, in order to address the problems of DC and uPIT, we propose a discriminative learning method for speaker-independent speech separation with deep embedding features. uPIT is incorporated into DC-based speech separation framework. Firstly, a DC network is trained to extract deep embedding features. Clusters in the embedding space can represent the inferred spectral masking patterns of individual source. Therefore, these deep embedding features contain the information of each source, which is conducive to speech separation. Then instead of using K-means clustering to estimate hard masks, we make full use of the uPIT network to directly learn each source's soft mask from the embedding features. And then uPIT and DC are jointly trained, which directly optimizes the actual separation objectives. Finally, in order to decrease the possibility of remixing the separated sources, motivated by our previous work [10], we apply the discriminative learning to fine tuning the separated model.

The rest of this paper is organized as follows. Section 2 presents the single channel speech separation based on masks. The proposed method is stated in section 3. Section 4 shows detailed experiments and results. Section 5 draws conclusions.

http://dx.doi.org/10.21437/Interspeech.2019-1940

2. Single Channel Speech Separation

The object of single channel speech separation is to separate target sources from a mixed signal.

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

where $y(t)$ is the mixed speech, S is the number of source signals and $x_s(t)$, $s = 1, ..., S$ are target sources. The corresponding short-time Fourier transformation (STFT) of those signals are $Y(t, f)$ and $X_s(t, f)$.

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 [18]. According to the commonly used masking method, the estimated magnitude $|\widetilde{X}_s(t, f)|$ of each source can be estimated by

$$|\widetilde{X}_s(t, f)| = |Y(t, f)| \odot M_s(t, f) \tag{2}$$

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 signals are utilized to reconstruct each source signal by inverse STFT (ISTFT).

3. The Proposed Speech Separation System

In this section, we present our proposed discriminative learning method for speaker-independent speech separation with deep embedding features, which is shown in Figure 1. From DC network [12] we can know that clusters in the deep embedding space can represent the inferred spectral masking patterns of individual sources. Therefore, these embedding vectors are discriminative features for speech separation. Motivated by this, we use this deep embedding vectors as the input of separation system. Then a uPIT network is used to learn the soft mask of each source instead of K-means clustering. Moreover, in oder to maximize the distance of each speaker, the discriminative learning is applied to fine tuning the whole model. Finally, uPIT and DC are jointly optimized.

3.1. Deep embedding features

As shown in Figure 1, we firstly train a BLSTM network as the extractor to acquire deep embedding features. The input of BLSTM is the mixed amplitude spectrum $|Y(t, f)|$ and the output is the D-dimensional deep embedding features V.

$$V = f_\theta(|Y(t, f)|) \in \mathbb{R}^{TF \times D} \tag{3}$$

where TF is the number of T-F bins and $f_\theta(*)$ is a mapping function based on the BLSTM network.

The loss function of embedding features' network is defined as follow:

$$\begin{aligned} J_{DC} &= ||VV^T - BB^T||_F^2 \\ &= ||VV^T||_F^2 - 2||V^TB||_F^2 + ||BB^T||_F^2 \end{aligned} \tag{4}$$

where $B \in \mathbb{R}^{TF \times C}$ is the source membership function for each T-F bin, i.e.,$B_{tf,c} = 1$, if source c has the highest energy at time t and frequency f compared to the other sources. Otherwise, $B_{tf,c} = 0$. C is the number of source. $|| * ||_F^2$ is the squared Frobenius norm. J_{DC} is the DC loss in Figure 1.

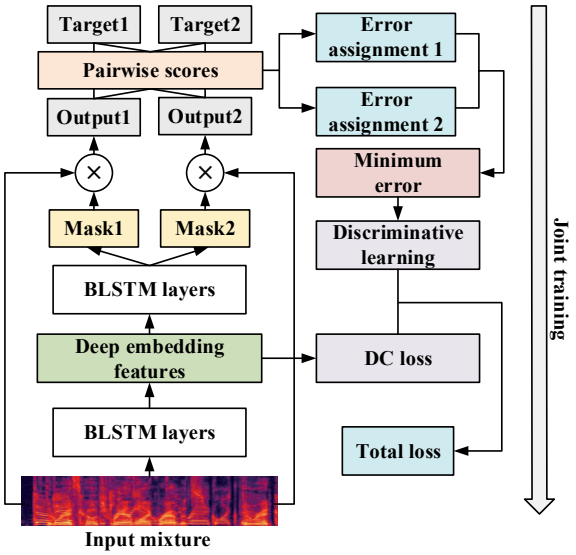

Figure 1: *Schematic diagram of our proposed DL-DEF speech separation system. DC loss is the loss of deep clustering.*

3.2. Speech separation model based on deep embedding features

Different from deep clustering [12] utilizing the K-means clustering to acquire hard masks, we use the embedding vectors as the input of uPIT to directly learn each source's soft masks. Therefore, the DC and uPIT can be trained end-to-end. When these features are extracted, they are reshaped as $V' \in \mathbb{R}^{T \times FD}$. Then these embedding features V' are sent to the separated system. In this way, we can learn a soft mask of each source from these features, which is better than the hard mask estimated by K-means clustering.

In order to make full use of phase information, we use the ideal phase sensitive mask (IPSM) [19] in this work. The IPSM is defined as

$$M_s(t, f) = \frac{|X_s(t, f)|cos(\theta_y(t, f) - \theta_s(t, f))}{|Y(t, f)|} \tag{5}$$

where $\theta_y(t, f)$ and $\theta_s(t, f)$ are the phase of mixed speech and target source.

During the separation stage, the uPIT is used to estimate each source. We apply the MSE between estimated magnitude and true magnitude as the training criterion.

$$J_{\phi_p(s)} = \frac{1}{TF} \sum_{s=1}^{S} |||Y| \odot \widetilde{M}_s - |X_{\phi_p(s)}|cos(\theta_y - \theta_s)||_F^2 \tag{6}$$

where $\widetilde{M}_s$ is the estimated mask. $\phi_p(s), p \in [1, P]$ is an assignment of target source s, $P = S!$ (! denotes the factorial symbol) is the number of permutations.

In order to solve the label ambiguity problem, the minimum cost among all permutations (P) is chosen as the optimal assignment.

$$\phi^* = \underset{\phi \in P}{arg\,min}\, J_{\phi_p(s)} \tag{7}$$

3.3. Discriminative learning

For uPIT, the target of minimizing Eq.7 is to make the predictions and targets more similar. In this paper, we explore dis-

criminative objective function that not only increase the similarity between the prediction and its target, but also decrease the similarity between the prediction and the interference sources.

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

$$J_{DL} = \phi^* - \sum_{\phi \neq \phi^*, \phi \in P} \alpha\phi \tag{8}$$

where ϕ is a permutation from P but not ϕ^*, $\alpha \geq 0$ is the regularization parameter of ϕ. When $\alpha = 0$, there is no discriminative learning, which is same as the loss of uPIT.

For two-talker speech separation, we assume that ϕ_1 is the permutation with the lowest MSE. Therefore, the cost function becomes as follow:

$$\begin{aligned} J_{DL} &= \phi_1 - \alpha\phi_2 \\ &= \frac{1}{TF}\sum(|||Y| \odot \widetilde{M}_1 - |X_1|||_F^2 - \alpha|||Y| \odot \widetilde{M}_1 - |X_2|||_F^2 \\ &+ |||Y| \odot \widetilde{M}_2 - |X_2|||_F^2 - \alpha|||Y| \odot \widetilde{M}_2 - |X_1|||_F^2) \end{aligned} \tag{9}$$

From Eq.9 we can know that the discriminative learning enlarges the distance of the target source with the interference sources. It means that it maximizes the differences between the target speakers with the others.

Therefore, the proposed model with discriminative learning minimizes the distance between the outputs of model and their corresponding reference signals. Simultaneously, it maximizes the dissimilarity between the target source and the interference. So the discriminative learning decreases the possibility of remixing the separated sources.

3.4. Joint training loss function

The deep clustering objective can reduce within-source variance in the internal representation [17]. Therefore, in order to effectively extract embedding features, we make full use of a joint training framework for the proposed system. More specifically, the deep clustering objective is added at the loss function.

$$\begin{aligned} J &= \lambda J_{DC} + (1-\lambda) J_{DL} \\ &= \lambda J_{DC} + (1-\lambda)(\phi^* - \sum_{\phi \neq \phi^*, \phi \in P} \alpha\phi) \end{aligned} \tag{10}$$

where J is the joint training loss function of the proposed system. $\lambda \in [0, 1]$ controls the weight of two objectives. Note that when $\lambda = 1$, the proposed method is same as deep clustering [12].

In order to get the better deep embedding features, we only train the DC network firstly. The loss function is Eq.4. Then these deep embedding features are used to train the separated model based on uPIT. The loss function is with no discriminative learning:

$$J' = \lambda J_{DC} + (1-\lambda)\phi^* \tag{11}$$

Finally, we apply the discriminative learning to fine tuning the whole model by the joint training loss function J in Eq.10.

4. Experiments and Results

4.1. Dataset

Our experiments are conducted on the WSJ0-2mix dataset [12], which is derived from WSJ corpus [20]. The WSJ0-2mix dataset consists three sets: training set (20,000 utterances about 30 hours), validation set (5,000 utterances about 10 hours) and test set (3,000 utterances about 5 hours). Specifically, training and validation set are generated by randomly selecting utterances from WSJ0 training set (`si_tr_s`) with signal-to-noise ratios (SNRs) between 0dB and 5dB. Similar as generating training and validation set, the test set is created by mixing the utterances from the WSJ0 development set (`si_dt_05`) and evaluation set (`si_et_05`).

We use the validation set 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 normalized 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. Our models are implemented using Tensorflow deep learning framework [21].

In this work, the deep embedding network has two BLSTM layers with 896 units. The embedding dimension D is set to 40. A tanh activation function is followed by the embedding layer. As for the separated network, it has only one BLSTM layer with 896 units. Therefore, there are three BLSTM layers in this work, which keeps the network configuration the same as baseline in [15]. A Rectified Liner Uint (ReLU) activation function is followed by the uPIT network, which is the mask estimation layer. The regularization parameter α of discriminative learning is set to 0.1.

All models contain random dropouts with a dropout rate 0.5. Each minibatch contains 16 randomly selected utterances. The minimum number of epoch is 30. The learning rate is initialized as 0.0005 and scaled down by 0.7 when the training objective function value increased on the development set. The early stopping criterion is that the relative loss improvement is lower than 0.01. Our models are optimized with the Adam algorithm [22].

4.3. Baseline model and evaluation metrics

We re-implement uPIT with our experimental setup as our baseline. It has three BLSTM layers with 896 units. The others are same as our experimental setup. In this work, in order to quantitatively evaluate speech separation results, the models are evaluated on the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) and signal-to-artifact ratio (SAR) which are the BBS-eval [23] score, and the perceptual evaluation of speech quality (PESQ) [24] measure.

4.4. Experimental results

Table 1 shows the results of SDR, SIR, SAR and PESQ between the proposed method and uPIT-BLSTM on the WSJ0-2mix database. DEF denotes the deep embedding features.

4.4.1. Evaluation of deep embedding features

From Table 1, we can find that our proposed uPIT+DEF methods outperform baseline uPIT in all objective measures no matter optimal assignment (Opt. assign.) or default assignment

Table 1: *The results of SDR, SIR, SAR and PESQ for different separation methods on closed (CC) and open (OC) condition. λ is the weight of joint training in Eq.10 and 11. DEF denotes the deep embedding features. uPIT is the baseline method, uPIT+DEF and uPIT+DEF+DL are our proposed methods. uPIT+DEF means with no discriminative learning.*

Method	λ	Optimal (Opt.) Assign.								Default (Def.) Assign.							
		SDR(dB)		SIR(dB)		SAR(dB)		PESQ		SDR(dB)		SIR(dB)		SAR(dB)		PESQ	
		CC	OC	CC	OC	CC	OC	CC	OC	CC	OC	CC	OC	CC	OC	CC	OC
uPIT(baseline)	-	11.3	11.2	18.8	18.8	12.3	12.3	2.68	2.67	10.3	10.1	17.7	17.5	11.5	11.3	2.60	2.58
uPIT+DEF	0.01	11.7	11.6	19.4	19.5	12.7	12.6	**2.85**	**2.84**	10.8	10.7	18.4	18.4	**12.0**	11.8	**2.77**	2.75
uPIT+DEF	0.05	11.7	11.7	19.5	19.6	12.7	12.6	2.84	**2.84**	10.8	**10.8**	18.4	**18.8**	11.9	**11.9**	2.76	2.75
uPIT+DEF	0.1	11.7	11.7	19.5	19.5	12.7	12.6	2.84	**2.84**	10.8	10.7	18.5	18.4	**12.0**	**11.9**	2.76	2.74
uPIT+DEF+DL	0.05	**11.9**	**11.9**	**19.9**	**20.0**	**12.8**	**12.7**	2.83	2.83	**11.0**	**10.8**	**18.8**	**18.8**	**12.0**	**11.9**	2.74	2.73

(Def. assign.). These indicate that the deep embedding features are more easily separated than the mixed amplitude spectral features. Because deep embedding features contain the potential masks of individual sources and they can effectively discriminate different target speakers. Therefore, they are conducive to speech separation.

Moreover, in order to acquire better deep embedding features, we propose a novel joint training framework to instruct the training of deep embedding network. Three different weights λ (0.01, 0.05 and 0.1) are applied. From Table 1, we can know that the performance of these three λ for speech separation are similar. The reason is that we firstly train a DC network with 30 epochs, which can obtain a pretty good representation for deep embedding features. Therefore, these three λ get similar performance.

4.4.2. Evaluation of discriminative learning

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 interferences. Therefore, we use discriminative learning to fine tuning the whole model based on $\lambda = 0.05$. From Table 1, we can know that when the discriminative learning is applied, our proposed method uPIT+DEF+DL achieves better performances than the proposed uPIT+DEF overall objective scores, except for the PESQ measure. This shows the effectiveness of the discriminative learning. Meanwhile, compared with the uPIT baseline system, the proposed uPIT+DEF+DL gets a better performance in all cases. For example, as for the optimal assignment on open condition, the proposed method achieves 6.3%, 6.4% and 3.3% relative improvements in SDR, SIR and SAR over the uPIT baseline system. These results reveal the effectiveness of our proposed method.

4.4.3. Comparisons with other separation methods

In order to better compare the performance of our proposed method (uPIT+DEF+DL) and other separation methods, Table 2 presents the results of SDR (dB) in the other competitive approaches on the same WSJ0-2mix dataset. Note that, for [9, 12, 25, 15, 26, 13] methods are use SDR improvements results. Therefore, we manually add 0.2 dB to their final results although the SDR result of the mixture is only about 0.15 dB. Compared with other speech separation methods, our proposed method improves the performance significantly to 11.9 dB and 10.8 dB with no phase enhancement for Opt Assign and Def Assign. Moreover, from Table 2, we can know that our proposed method outperforms other methods, such as DC+, DANet, SDC-MLT-Grid LSTM. Compared with DC+ [26], our proposed method achieves 12.5% relative improvement on open condition. These results confirm that using discriminative learning and deep embedding features can improve the performance of speaker-independent speech separation.

Table 2: *The results of SDR (dB) in the other different separation methods on the WSJ0-2mix dataset on CC and OC with no phase enhancement.*

Method	Opt Assign		Def Assign	
	CC	OC	CC	OC
DC[12]	-	-	6.1	6.0
DC+[26]	-	-	-	9.6
DANet[13]	-	-	-	9.8
uPIT-BLSTM [15]	11.1	11.0	9.6	9.6
cuPIT-Grid LSTM-RD[9]	11.4	11.4	10.4	10.3
SDC-MLT-Grid LSTM [25]	11.6	11.6	10.8	10.7
uPIT+DEF+DL(our proposed)	**11.9**	**11.9**	**11.0**	**10.8**

5. Conclusions

In this paper, we propose a speaker-independent speech separation method with discriminative learning based on deep embedding features. We firstly train an DC network to extract deep embedding features. Then these features are used as the input of uPIT system to directly separate the different speaker sources. Moreover, uPIT and DC are jointly optimized. Finally, the discriminative learning is applied to fine tuning the whole model. Results show that the proposed method outperforms uPIT baseline, with a relative improvement of 6.3%, 6.4% and 3.3% relative improvements in SDR, SIR and SAR, respectively. In the future, we will explore phase enhancement based on the proposed method.

6. Acknowledgements

This work is supported by the National Key Research & Development Plan of China (No.2017YFB1002802), the NSFC (No.61425017, No.61831022, No.61771472, No.61603390), the Strategic Priority Research Program of Chinese Academy of Sciences (No.XDC02050100), and Inria-CAS Joint Research Project (No.173211KYSB20170061). Authors also thank Shuai Nie for his helpful comments on this work.

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] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[3] 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.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[5] 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.

[6] H. Erdogan and T. Yoshioka, "Investigations on data augmentation and loss functions for deep learning based speech-background separation," *Proc. Interspeech 2018*, pp. 3499–3503, 2018.

[7] J. Wang, J. Chen, D. Su, L. Chen, M. Yu, Y. Qian, and D. Yu, "Deep extractor network for target speaker recovery from single channel speech mixtures," *Proc. Interspeech 2018*, pp. 307–311, 2018.

[8] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *Proc. ICASSP*. IEEE, 2018, pp. 696–700.

[9] C. Xu, W. Rao, X. Xiao, E. S. Chng, and H. Li, "Single channel speech separation with constrained utterance level permutation invariant training using grid lstm," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6–10.

[10] C. Fan, B. Liu, J. Tao, Z. Wen, J. Yi, and Y. Bai, "Utterance-level permutation invariant training with discriminative learning for single channel speech separation," in *Proc. ISCSLP*. IEEE, 2018.

[11] K. Wang, F. Song, and X. Lei, "A pitch-aware approach to single-channel speech separation," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[12] 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.

[13] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 246–250.

[14] D. Yu, M. Kolbk, 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.

[15] M. Kolbæk, D. Yu, Z. Tan, J. Jensen, M. Kolbaek, D. Yu, Z. 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.

[16] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 686–690.

[17] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 61–65.

[18] 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.

[19] 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.

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

[21] 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.

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

[23] 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.

[24] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part i–time-delay compensation," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 755–764, 2002.

[25] C. Xu, W. Rao, E. S. Chng, and H. Li, "A shifted delta coefficient objective for monaural speech separation using multi-task learning," in *Proceedings of Interspeech*, 2018, pp. 3479–3483.

[26] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.