CBLDNN-BASED SPEAKER-INDEPENDENT SPEECH SEPARATION VIA GENERATIVE ADVERSARIAL TRAINING

Chenxing Li[1,2]*, Lei Zhu*[3]*, Shuang Xu*[1]*, Peng Gao*[3]*, Bo Xu*[1]

[1]Institute of Automation, Chinese Academy of Sciences, Beijing, P.R.China
[2]University of Chinese Academy of Sciences, Beijing, P.R.China
[3]AI Lab, Rokid Inc., Beijing, P.R.China

ABSTRACT

In this paper, we propose a speaker-independent multi-speaker monaural speech separation system (CBLDNN-GAT) based on convolutional, bidirectional long short-term memory, deep feed-forward neural network (CBLDNN) with generative adversarial training (GAT). Our system aims at obtaining better speech quality instead of only minimizing a mean square error (MSE). In the initial phase, we utilize log-mel filterbank and pitch features to warm up our CBLDNN in a multi-task manner. Thus, the information that contributes to separating speech and improving speech quality is integrated into the model. We execute GAT throughout the training, which makes the separated speech indistinguishable from the real one. We evaluate CBLDNN-GAT on WSJ0-2mix dataset. The experimental results show that the proposed model achieves 11.0dB signal-to-distortion ratio (SDR) improvement, which is the new state-of-the-art result.

Index Terms— speech separation, multi-task learning, generative adversarial training, CBLDNN.

1. INTRODUCTION

In recent years, speaker-independent multi-speaker separation has attracted more and more attention. The purpose of this issue is to separate the mixed speech of multiple speakers into the speech of each person, which improves the performance of conference transcription system, human-computer interaction system and large vocabulary continuous speech recognition (LVCSR) system.

Many approaches have been proposed for decades. Computational auditory scene analysis (CASA) [1, 2] is widely adopted. Based on perceptual grouping cues, it cuts mixed speech into time-frequency (T-F) segments, which are assumed to be derived from the different sources. Independent streams are generated from these segments by clustering. Non-negative matrix factorization (NMF) based methods [3, 4, 5] decompose signal into sets of bases and weight matrices, which are used to estimate mixing factors during evaluation. Both CASA and NMF have limited performance. GMM-HMM based separations [6, 7] also show impressive results, but these methods only work well in close-set speaker condition.

Deep neural network is gradually applied to speech separation. Deep clustering (DPCL) [8] achieves impressive results. A DNN is trained to derive embeddings for each T-F bin to optimize segmentation criterion. During evaluation, each T-F bin is mapped into embedding space. Clustering is then applied to generating partition. Deep attractor is proposed in [9], where 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. A novel training criterion named permutation invariant training (PIT) is proposed in [10, 11]. PIT determines the best output assignment automatically and then minimizes the error given the assignment, which is implemented inside the network. It solves the label permutation problem and integrates speaker tracing in PIT. Thus separation and tracing can be trained in one step.

However, the methods above come with several shortcomings. For example, researchers assume that each T-F bin only belongs to one speaker in DPCL, which is sub-optimal. It is also inefficient to perform end-to-end mapping. The limitation of deep attractor is the requirement to estimate attractor points during test. Besides, the MSE loss in all methods only concerns the numerical difference in the estimation, and the numerical error reduction may not necessarily lead to perceptual improvement on the separated speech. Similarly, in super-resolution (SR) problems [12], texture detail in the reconstructed SR images is typically absent after optimizing MSE loss. The same in our task, the ability of MSE to capture perceptually relevant differences, such as high texture detail, is very limited as they are defined based on each T-F differences.

Short time fourier transform (STFT) spectral magnitude contains sufficient information and has the advantage of fast calculation, which is used as input. Log-mel filterbank (fBank) is based on human hearing perceptions. Pitch is selected as an important cue in CASA [2]. In order to enable the model to take advantage of these information to make separated speech has a better quality and use these features to drive the network into a convergent state because of their compact structure. fBank-pitch feature is fed into training by multi-task learning (MTL) [13]. Our contributions are as follows: (1). A more sophisticated structure, CBLDNN, is used to improve the performance; (2). MTL strategy is adopted to warm up the network for the first 10 epochs; (3). GAT is employed to train the network to further enhance the speech quality.

The experiments are conducted on WSJ0-2mix dataset [8], and SDR [14] improvement is utilized to evaluate the performance. Compared with uPIT-BLSTM in paper [10], experimental results show that our CBLDNN-based models achieve better performance when using the same training criterion. Through 10 rounds of MTL, an average increase of 0.25dB SDR improvement is obtained. After GAT, the proposed system, CBLDNN-GAT, achieves 11.0dB SDR improvement, which outperforms the current algorithms.

This paper is organized as follow. Section 2 introduces monaural speech separation. Section 3 describes GAT used in this paper. Experimental setup and results of our experiment are presented in Section 4. Finally, Section 5 concludes our work.

This work is supported by National Key Research and Development Program of China under Grant No.2016YFB1001404 and Beijing Engineering Research Center Program under Grant No.Z171100002217015

978-1-5386-4658-8/18/$31.00 ©2018 IEEE

2. MONAURAL SPEECH SEPARATION AND TRAINING TARGETS

Speech separation aims at estimating individual source signals in mixed speech. In this paper, we focus on monaural speech separation task. The source signals are assumed linearly mixed, which can be represented as:

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

where S is the number of source signals. $x_s(n)$ and $y(n)$ denote the s-th source signal and the mixed speech, respectively. The following relationship is still satisfied after STFT

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

where $Y(t,f)$ and $X_s(t,f)$ represent the STFT of speech $y(n)$ and $x_s(n)$ respectively. Thus, our task is clarified as recovering each source signal $x_s(n)$ from $y(n)$ or $Y(t,f)$. It is well-known that better results can be obtained by estimating a set of masks [15]. In our experiment, we firstly deploy a deep neural network to estimate a set of masks $M(t,f)$ in frequency domain instead of directly recovering $x_s(n)$ from $y(n)$. In the following equation, $H_s(|Y(t,f)|,\theta)$ represents a non-linear representation from STFT spectral magnitude $|Y(t,f)|$ to $M_s(t,f)$. $M_s(t,f)$ represents the mask of the s-th signal

$$H_s(|Y(t,f)|,\theta) = M_s(t,f), \tag{3}$$

and $|X_s(t,f)|$ can be recovered by $M_s(t,f) \times |Y(t,f)|$ (× indicates element-wise multiplication). The separated speech signal $x_s(n)$ can be obtained after inverse STFT.

Masks are to be estimated as the training targets, and three widely-accepted masks [10, 16] are utilized in this paper. The ideal ratio mask (IRM) for each source is defined as

$$M_s^{IRM} = |X_s(t,f)| / \sum_{s=1}^{S} |X_s(t,f)|. \tag{4}$$

The IRM maximizes the SDR when the phase of $y(n)$ is used for reconstruction and all sources have the same phase, which is an invalid assumption. Besides, $|X_s(t,f)|$ can not be obtained in practice. Like [10], IRM is used to compute an upper bound of performance. Another mask is ideal amplitude mask (IAM), which for each source is defined as

$$M_s^{IAM} = |X_s(t,f)| / |Y_s(t,f)|. \tag{5}$$

IAM can achieve the highest SDR when the phase of each source equals the mixed speech. Since IAM ignores the differences of phases, phase sensitive mask (PSM) is put forward to address this issue. The PSM for each source is defined as

$$M_s^{PSM} = |X_s(t,f)| \times cos(\theta_y(t,f) - \theta_s(t,f)) / |Y_s(t,f)|, \tag{6}$$

where $\theta_y(t,f)$ is the phase of mixed speech, and $\theta_s(t,f)$ is the phase of s-th source signal.

3. GENERATIVE ADVERSARIAL TRAINING

Generative adversarial net [17] comprises of two adversarial sub networks, a generator which generates the fake examples from the random noises, and a discriminator which discriminates whether the input is real or generated by the generator.

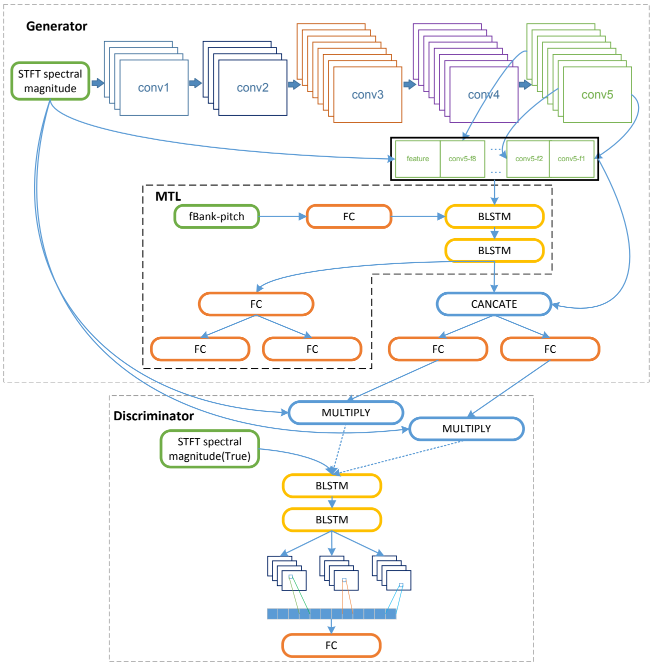


Fig. 1: CLDNN-based speech separation with GAT.

In this paper, we implement a conditional GAN [18, 19], where the generator, a CLDNN-based model, performs mapping conditioned on some extra information. Specifically, the generator learns a mapping from observed feature $|Y(t,f)|$ to each mask $M_s(t,f)$. The discriminator is trained to classify whether the STFT feature comes from real speech or generated. This training procedure is diagrammed in Fig. 1.

3.1. CBLDNN-based generator with MTL

For solving tracing and permutation problem, an utterance-level CLDNN-based generator is proposed. Convolutional layers are good at modeling frequency variations [20]. Some acoustic variations can be effectively normalized and the resultant feature representation may be immune to speaker variations, colored background and channel noises. Besides, filters that work on local frequency region provide an efficient way to represent local structures and their combinations, which give a more precise spectral structure to separated speech. BLSTMs well model temporal variations and DNN layers map features to a more separable space. The CBLDNN architecture incorporates the three layers in a unified framework, fusing the benefits of individual layers.

The proposed CBLDNN model is similar to [21], but with fine adjustment and more sophisticated structure. As depicted in Fig. 1. The proposed model consists of 5 stacked convolutional layers, 2 stacked BLSTM layers and 2 independent output layers. In (inChannel, outChannel, kernelW, kernelH) format, the convolutional layers have one (1,4,5,5), one (4,4,3,3), one (4,8,5,5), one (8,8,3,3) and one (8,8,1,1) convolution layer with no pooling and 1 stride in height and width. Each BLSTM layer has 400 units. The model has 2 fully-connected (FC) independent output layers and each has 129 output nodes. The output is the mask of separated speech.

MTL is an effective approach to improve the performance of a single task with the help of other related tasks [13]. It provides a convenient way of combining information from multiple tasks. The human perception of the frequency contents of speech signals is nonlinear. fBank is based on the human peripheral auditory system. Pitch is an important cue in CASA [2]. We aim to separate the speech not only with considerable numerical error reduction but with high quality. We use fBank-pitch based speech separation as another task in MTL. In this experiment, we only share BLSTM layers, and ReLU is adopted as activation function of output layer, which is depicted in Fig. 1. Besides, we only use MTL for 10 rounds, which is also used to warm up the network.

3.2. BLCDNN-based discriminator

In our experiment, we use utterance-level BLCDNN-based discriminator, which is depicted in Fig. 1.

We aim to use BLSTM to model the dependency of the speech and convolutional layer to extract discriminative features that are useful for classification task. In convolutional layer, to enable the network to extract complementary features and enrich the representation, we learn several different filters simultaneously. Convolutional filters with multiple sizes capture valuable features from different scales, which contribute a lot to robust classification. The feature maps produced by the convolution layer are forwarded to the pooling layer. 1-max pooling is employed on each feature map to reduce it to a single but the most dominant feature. The features are then joined to form a feature vector input to the final layer. This step transforms the variable-length, high-dimensional vector into a fixed-length vector. Finally, a fully-connected layer maps it to one output node. The input is more like a real speech when the output value is closer to 1.

The BLCDNN model consists of 2 BLSTM layers, 1 convolutional layer and 1 fully-connected layer. Each BLSTM layer has 256 units. In (kernelW, kernelH) format, the convolutional layer has 3 different filter sizes that are (5, 5), (3, 3) and (1, 1) both with 4 output channels and 1 stride in height and width.

3.3. Loss function

For comparison, we use utterance-level PIT-based speech separation systems [10] as baselines. For end-to-end training, we firstly restore $|\hat{X}(t, f)|$, which means STFT spectral magnitude generated by generator. The loss function based on IAM is

$$\mathcal{L}_2^{IAM} = \frac{1}{N}\sum_{s=1}^{S} ||M_s^{IAM} \times |Y| - |X_{\phi^*}|||^2, \tag{7}$$

where M, $|Y|$ and $|X|$ represent mask, STFT spectral magnitude of mixed speech and STFT spectral magnitude of source signal for one utterance respectively. N is the total number of T-F bins over all sources, and ϕ^* is the permutation that minimizes the utterance-level separation error defined as

$$\phi^* = \arg\min_{\phi} \sum_{s=1}^{S} ||M_s^{IAM} \times |Y| - |X_{\phi^*}|||^2. \tag{8}$$

For PSM-based method, the loss function can be modified as

$$\mathcal{L}_2^{PSM} = \frac{1}{N}\sum_{s=1}^{S} ||M_s^{PSM} \times |Y| - |X_{\phi^*}| \cos(\theta_y - \theta\phi(s))||^2, \tag{9}$$

where ϕ^* is the permutation that minimizes the utterance-level separation error defined as

$$\phi^* = \arg\min_{\phi} \sum_{s=1}^{S} ||M_s^{PSM} \times |Y| - |X_{\phi^*}| \cos(\theta_y - \theta\phi(s))||^2. \tag{10}$$

The equations above are traditional MSE-based PIT loss functions. In this experiment, we train the network by GAT. Thus the loss function is modified. We use LSGAN [22] based method. At the same time, L_1-regularization is utilized to guide the training. In order to balance GAN loss and L_1-regularization, λ is taken as hyper-parameter in this experiment

$$\begin{aligned} \min_D \mathcal{L}(D) &= \mathbb{E}_{|X|\sim p_{data}(|X|)}[(D(|X|) - 1)^2] \\ &\quad + \mathbb{E}_{|Y|\sim p_{data}(|Y|)}[(D(G(|Y|) \times |Y|))^2], \\ \min_G \mathcal{L}(G) &= \mathbb{E}_{|Y|\sim p_{data}(|Y|)}[(D(G(|Y|) \times |Y|) - 1)^2] + \lambda \mathcal{L}_1^{PSM}. \end{aligned} \tag{11}$$

After employing MTL, our final loss function has the following form

$$\mathcal{L} = \begin{cases} \mathcal{L}_{T_1} + \mu\mathcal{L}_{T_2}, & \textit{if epoch} \leq \textit{10} \\ \mathcal{L}_{T_1}, & \textit{if epoch} > \textit{10} \end{cases}, \tag{12}$$

where $\mathcal{L}$ is the total loss. $\mathcal{L}_{T_1}$ is the loss of $\mathcal{L}_2^{PSM}$ or $\mathcal{L}(G)$ based on the training method. $\mathcal{L}_{T_2}$ represents the loss of the second task, fBank-pitch based PIT. μ is a hyper-parameter used to balance the loss between multiple tasks.

4. EXPERIMENTS

4.1. Experimental setup

We evaluate CLDNN-GAT on WSJ0-2mix dataset [8], which is derived from WSJ corpus [23]. WSJ0-2mix contains 30 hours of training data, 10 hours of development data (Dev set) and 5 hours of test data (Test set).

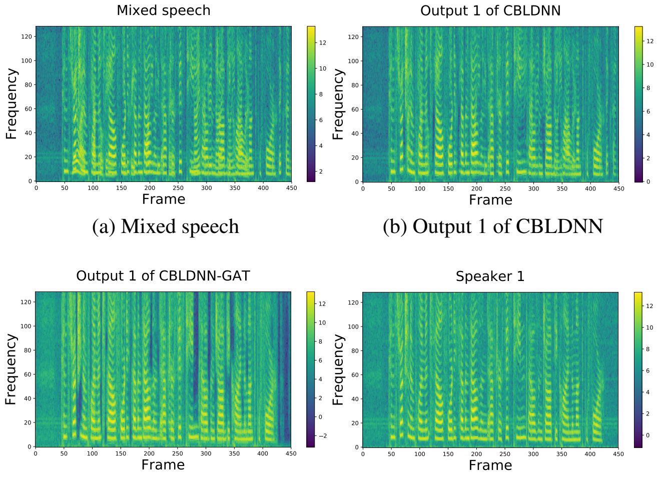
The input features of generator and discriminator are 129-dimensional STFT spectral magnitude computed with a frame size of 32ms and 16ms shift. For MTL, 40-dimensional fBank features and 3-dimensional pitch features are extracted. The phase of the source signal is used to build PSM-based loss function, and the phase of the mixed speech is used to restore the speech. After fine adjustment, hyper-parameters λ and μ are set as 1 and 0.5 respectively. fBank-pitch is extracted by Kaldi [24] and the models are all trained on Tensorflow [25]. RMSprop algorithm [26] is utilized for training where the learning rate started at 0.0002.

4.2. Baseline systems

In this experiment, we conduct several CBLDNN-based baseline systems by using utterance-level PIT [10]. The experimental results are shown in Table 1. From the table, we can see that the systems trained with phase information obtain better performance. This shows that phase information does improve performance. Among baselines, CBLDNN-PSM-ReLU obtains the best result, which improves SDR by 9.7dB in the test set. At the same time, compared with uPIT-BLSTM in [10], our baseline systems achieve better performance when using the same training method, about 0.45dB SDR improvement in average is obtained. This shows that CNN effectively extracts local features and applies them to subsequent separation. It also indicates that our CBLDNN systems have better modeling capabilities.

4.3. CBLDNN-based systems with MTL

CBLDNN-MTL is trained with extra MTL compared with CBLDNN in section 4.2. In this experiment, MTL is applied at the first 10 rounds. The experimental results are shown in Table 1.



(a) Mixed speech

(b) Output 1 of CBLDNN

(c) Output 1 of CBLDNN-GAT

(d) Source signal

Fig. 2: An example of separated speech and its source signal, which is randomly selected. The output of CBLDNN achieves 10.02dB SDR improvement, and the output of CBLDNN-GAT achieves 11.5dB SDR improvement.

Table 1: *SDR improvement for different separation methods.*

Model	Mask	Activation	SDR Imp. (dB)	
			Dev set	Test set
CBLDNN	IAM	Sigmoid	8.2	8.3
		ReLU	8.7	8.9
		Softmax	8.9	8.9
	PSM	Sigmoid	9.4	9.3
		ReLU	**9.6**	**9.7**
		Softmax	9.6	9.6
CBLDNN-MTL	IAM	Sigmoid	8.8	8.8
		ReLU	9.0	9.1
		Softmax	9.2	9.2
	PSM	Sigmoid	9.6	9.6
		ReLU	9.7	9.7
		Softmax	**9.8**	**9.8**

From Table 1 we can see that MTL can be used to improve SDR. Compared with baselines, an average increase of 0.25dB SDR improvement is obtained. CBLDNN-IAM-Sigmoid-MTL has achieved the most significant performance improvement compared with corresponding baseline system, which is 0.5dB SDR improvement. Besides, CBLDNN-PSM-Softmax-MTL achieves best SDR improvement, 9.8dB. This shows that the information of fBank-pitch has been modeled by network successfully, and this information can be used to improve the performance.

4.4. CBLDNN-based systems with GAT

Table 2: *SDR improvement for different separation methods.*

Model	Activation	SDR Imp. (dB)	
		Dev set	Test set
CBLDNN-L1 loss	Sigmoid	9.7	9.7
	ReLU	9.6	9.6
	Softmax	9.5	9.6
CBLDNN-GAT	Sigmoid	**11.0**	**11.0**
	ReLU	10.7	10.8
	Softmax	10.6	10.6
DPCL [8]	-	5.9	5.8
DPCL+ [9]	-	-	9.1
DPCL++ [27]	-	-	10.8
DANet [9]	-	-	9.6
DANet-6 anchor [28]	-	-	10.4
uPIT-BLSTM [10]	ReLU	9.4	9.4
uPIT-BLSTM-ST [10]	ReLU	10.0	10.0
IRM	-	12.4	12.7

In this section, we explore the impact of GAT. In the beginning, MTL is conducted to train the network for 10 rounds. GAT is applied to training the network from the beginning to the end. As phase information brings performance improvement, only PSM-based method is evaluated in this section. In GAT, PSM is used to measure the loss of L_1, which is utilized to minimize the distance between generations and the clean examples.

We adopt L_1-regularization to GAT. Thus we attempt to find out whether L_1-regularization has greatly improved performance and the results show that the systems perform similarly as baselines do. With GAT, SDR has been significantly improved, which means that GAT plays a more important role in improving the performance. Our system, CBLDNN-GAT with sigmoid activation function achieves best results, with 11.0dB SDR improvement. At the same time, compared with exited methods, experimental results show that our system obtains the best results. GAT makes the separated speech produced by generator approaches to real one. Compared with PIT, our goal is no longer to only reduce the numerical differences between the separated speech and target speech but to separate the speech with high speech quality. In practice, the network does not need discriminator network. Thus we can achieve better performance compared with PIT while having the same network structure.

Fig. 2 shows the spectrogram of separated speech based on different separation methods. Compared with speech separated by CBLDNN, the speech separated by CBLDNN-GAT has clearer spectrum structure and complete high-frequency spectrum. At the same time, the speech separated by CBLDNN-GAT has limited interference. But the speech separated by CBLDNN contains obviously interference. More examples of separated speech are provided at https://github.com/chenxinglili/SpeechSeparationExamples

5. CONCLUSION AND DISCUSSION

In this paper, we introduce CBLDNN-based speaker-independent speech separation system with GAT. Our results on two-speaker mixed speech separation task indicate that CBLDNN-GAT can achieve a new state-of-the-art performance. Additionally, CBLDNN-GAT effectively deals with the label permutation and tracing problem. We note that the proposed method has great potential for the further improvement. Firstly, we can increase the number of training epochs of MTL to see if further improvement can be obtained. Secondly, we will extend the experiment to three-speaker mix task. Thirdly, from Fig. 2, the spectrums miss in some T-F bins. Although the missing parts have no voice, we will test the speech recognition performance to explore the impact and evaluate the speech quality in objective standards, such as perceptual evaluation of speech quality.

6. REFERENCES

[1] M. Cooke, *Modelling auditory processing and organisation*. Cambridge University Press, 2005, vol. 7.

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

[3] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *ISCA International Conference on Spoken Language Proceesing*, 2006.

[4] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[5] J. Le Roux, F. J. Weninger, and J. R. Hershey, "Sparse nmf–half-baked or well done?" *Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023*, 2015.

[6] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2299–2310, 2007.

[7] K. Hu and D. Wang, "An iterative model-based approach to cochannel speech separation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, p. 14, 2013.

[8] J. R. Hershey, Z. Chen, J. Le 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.

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

[10] 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 Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

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

[12] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv preprint arXiv:1609.04802*, 2016.

[13] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.

[14] 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.

[15] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1849–1858, 2014.

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

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv:1611.07004*, 2016.

[19] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[20] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 4277–4280.

[21] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4580–4584.

[22] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," *arXiv preprint ArXiv:1611.04076*, 2016.

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

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584, 2011.

[25] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[26] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[27] 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.

[28] Z. Chen, Y. Luo, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *arXiv preprint arXiv:1707.03634*, 2017.