A comprehensive study of speech separation: spectrogram vs waveform separation

Fahimeh Bahmaninezhad[1] [§], *Jian Wu*[2,4] [§], *Rongzhi Gu*[3,4]*, Shi-Xiong Zhang*[5]*, Yong Xu*[5]*, Meng Yu*[5]*, Dong Yu*[5]

[1]University of Texas at Dallas, Richardson TX, USA
[2]Northwestern Polytechnical University, Xi'an, China
[3]Peking University Shenzhen Graduate School, Shenzhen, China
[4]Tencent AI Lab, Shenzhen, China
[5]Tencent AI Lab, Bellevue WA, USA

`fahimeh.bahmaninezhad@utdallas.edu, ,{auszhang,lucayongxu,raymondmyu,dyu}@tencent.com`

Abstract

Speech separation has been studied widely for single-channel close-talk microphone recordings over the past few years; developed solutions are mostly in frequency-domain. Recently, a raw audio waveform separation network (TasNet) is introduced for single-channel data, with achieving high Si-SNR (scale-invariant source-to-noise ratio) and SDR (source-to-distortion ratio) comparing against the state-of-the-art solution in frequency-domain. In this study, we incorporate effective components of the TasNet into a frequency-domain separation method. We compare both for alternative scenarios. We introduce a solution for directly optimizing the separation criterion in frequency-domain networks. In addition to speech separation objective and subjective measurements, we evaluate the separation performance on a speech recognition task as well. We study the speech separation problem for far-field data (more similar to naturalistic audio streams) and develop multi-channel solutions for both frequency and time-domain separators with utilizing spectral, spatial and speaker location information. For our experiments, we simulated multi-channel spatialized reverberate WSJ0-2mix dataset. Our experimental results show that spectrogram separation can achieve competitive performance with better network design. Multi-channel framework as well is shown to improve the single-channel performance relatively up to +35.5% and +46% in terms of WER and SDR, respectively.

Index Terms: Speech separation, multi-channel separation framework, target speaker extraction, Si-SNR loss function.

1. Introduction

Speech separation refers to the task of extracting all overlapping speech sources in a given mix speech signal [1]. With growing speech activated technologies, such as Amazon Alexa and Apple Siri, the cocktail party problem [2] has became more of an interest. Generally speaking, speech separation can be considered as a pre-processing step for other speech applications (e.g., speech and speaker recognition) when data has been collected with overlapping speech sources [3].

Source and speech separation has been studied for many years [4, 5, 6, 7, 8, 9]; however, after introducing deep clustering (DC) [10] approach for speech separation and generally deep learning based methods [1, 11], automatic speech separation performance is improving faster.

Most of speech separation techniques are studied in frequency-domain [1, 10, 12, 11, 13] for close-talk single-channel recordings; and over the past few years new network structures [14] or loss functions [12, 11] are introduced to increase the accuracy of these systems. In addition, authors in [14] proposed an end-to-end framework with incorporating STFT operation in network layers and introduced unfolded iterative phase reconstruction to further improve the performance (inspired by the effectiveness of phase information in speech separation objective criteria). Speech separation with effective deep learning approaches are also started to develop for multi-channel data as well [3, 15]. After all, recently [16] introduced a new solution, i.e., TasNet, for speech separation but in time-domain, which achieved an interesting performance comparing against the frequency-domain solutions.

In this paper, we investigate two approaches of speech separation; i.e., spectrogram (frequency-domain) and waveform (time-domain) separation in detail. We incorporate effective components of the TasNet into a spectrogram separation framework, such as: CNN-based separation network as well as loss function. We introduce an end-to-end spectrogram separation network with directly optimizing the Si-SNR criterion. We study both waveform and spectrogram separations for multi-channel data as well, with introducing IPD (inter-microphone phase difference) features [14]. We also evaluate the effectiveness of our multi-channel framework with speaker directional features, i.e., angle features, for extracting speech of a specific target speaker from the mix signal. Eventually, in addition to speech separation objective and subjective criteria, we report WER as well; considering the fact that speech separation is introduced to benefit other speech related applications, including speech recognition.

2. Problem Setup

In this section, we overview speech separation task in time and frequency-domain, introduce data used here for our experiments and describe our evaluation metrics.

2.1. Speech Separation: Spectrogram vs Waveform

In this section, we briefly describe spectrogram and waveform separation methods based on masking approaches. Our focus in this paper is also on masking based separation.

The methods performing in frequency-domain, first analyze the given mix speech signals into magnitude and phase (with STFT - short-time Fourier transform), then separate the magnitude part into S components (where S is the number of over-

---

4574   `http://dx.doi.org/10.21437/Interspeech.2019-3181`

lapping individual sources in the mix signal) with the separation network. For reconstructing the separated speech signals in time-domain, the input mix signal phase along with the separated magnitudes are fed into ISTFT (inverse STFT). In detail, for a mix signal y, the task is separating overlapping sources x_s such as $y = \sum_{s=1:S}(x_s)$. If we represent $|Y|$ and $\angle Y$ as magnitude and phase of signal y, respectively; and M as the mask functions learned with separation network, then for each individual source x_s the estimated magnitude is $|\hat{X}_s| = M_s \odot |Y_s|$ (where $\odot$ is element-wise multiplication). Separated signals are then reconstructed with ISTFT using inputs $|\hat{X}_s|$ and $\angle Y$. In [14] authors proposed iterative phase reconstruction algorithm which has shown to perform better than using the $\angle Y$ for reconstruction of the signal. In this paper, by lower-case letters, e.g., y, we mean time-domain samples, upper-letter Y frequency-domain representation and $|Y|$ as the magnitude; y_0 also means time-domain samples for channel 0.

On the other hand, the TasNet as a time-domain speech separation replaces the STFT and ISTFT components with trainable encoder and decoder, respectively. Therefore, instead of mapping samples into the frequency-domain, it maps them into a new temporal resolution space. The encoder does not analyze time samples into magnitude and phase components, it operates on a unified representation which makes it different from the spectrogram separators. In sum, the time-domain signal is mapped to a new domain with encoder and fed into the separation network. The separation network learns S mask functions which are multiplied by the output of the encoder; the S estimated separated sources are then converted back to the time-domain with decoder layer.

Throughout this paper, we use uPIT-BLSTM [1] as our baseline system for spectrogram separation; and TasNet [16] for waveform separation.

2.2. Data

We simulated a spatialized reverberant dataset derived from Wall Street Journal 0 (WSJ0)-2mix corpus, which is a well-studied dataset in monaural and multi-channel speech separation. There are 20,000, 5,000 and 3,000 multi-channel, reverberant, two speaker mixed speech in training, development and test subsets, respectively. The performance evaluation is all done on the test set, where all the speakers are unseen during training. We consider a 6-microphone circular array of 7-cm diameter, where speakers and the microphone-array are randomly located in the room. The two speakers and microphone array are on the same plane and all of them are at least 0.3m away from the wall. The image method is employed to simulate RIRs randomly from 3000 different room configurations with the size (length-width-height) ranging from 3m-3m-2.5m to 8m-10m-6m. The reverberation time T60 is sampled in a range of 0.05s to 0.5s. Samples with angle difference of 0-15°, 15-45°, 45-90°and 90-180°respectively account for 16%, 29%, 26% and 29% in the dataset. From these 6-channels, we always use the first channel as input to the network representing the mix signal.

2.3. Metric

Different metrics are introduced for evaluation of speech separation systems [17]. We use Si-SNR and SDR to objectively measure the separation accuracy.

$$\text{SDR} = 10\log_{10}\frac{||x_{target}||^2}{||e_{inter} + e_{noise} + e_{artif}||^2}, \qquad (1)$$

$$\text{Si-SNR} = 10\log_{10}\frac{||x_{target}||^2}{||e_{noise}||^2}, \qquad (2)$$

where $x_{target} = \frac{<\hat{x},x>x}{||x||^2}$. $\hat{x}$ and x respectively represent estimated and reference signals. For Si-SNR, the scale invariant is guaranteed by mean normalization of estimated and reference signals to zero mean [16].

Quality of separated speech sources is also evaluated with PESQ [18]. We also report WER (word error rate) to measure how the the alternative separation techniques can be effective in speech recognition tasks.

3. Improved Spectrogram Separation

Traditional TF-masking (time-frequency masking)/spectrogram separation networks are different than TasNet in three main components: 1) *Encode-decoder*: for spectrogram separation STFT/ISTFT is used to transform/reconstruct data into/from the frequency-domain, but in waveform separation, Conv-1d and ConvTranspose-1d are used respectively to encode and decode the data. Therefore, the latter one does not decompose the signal in magnitude and phase components (part of the gain achieved from waveform separation can be related to this fact). 2) *Separation network*: the separation networks used in [1] and [16] are also different. In [16], authors mentioned the CNN-based structure outperforms the BLSTM-based network for their task (although, for spectrogram separation, alternative network structures are studied besides the BLSTM layers [1, 14] 3) *Training loss*: the loss function used for TF spectrogram masking is usually MSE (mean square error) [1] which directly does not optimize the separation criterion. However, TasNet network is trained with Si-SNR loss function which directly optimizes the separation performance.

In this paper, we incorporate the CNN-based separation structure of TasNet into a spectrogram separation framework to examine the effectiveness of the network structure. In addition, we formulate the Si-SNR loss function for spectrogram separation as well; by replacing STFT and ISTFT with Conv-1d and ConvTranspose-1d using a specific kernel. Details of these two modifications are explained in the following two subsections.

3.1. CNN- and BLSTM-Based Separation Network

We use the BLSTM structure in [1] as baseline and evaluate the performance of the system when n×BLSTM layers are replaced with CNN-based separation network introduced in [16] (please refer to the original paper for details on the structure).

3.2. Si-SNR Loss Function

The traditional uPIT (utterance level permutation invariant training) loss function uses MSE to train the model parameters; and in other TF-masking methods as well variations of MSE have been chosen as the training loss. To the best of our knowledge Si-SNR loss was not used with frequency domain frameworks before. [14] implemented the STFT throughout the network layers but used phase sensitive MSE as their loss function.

At the end, speech separation performance will be reported in terms of Si-SNR and SDR. The TasNet network directly optimizes the Si-SNR value. Here, with some modifications in traditional implementation of STFT and ISTFT, we propose to employ Si-SNR loss while training the spectrogram separation.

Short-time Fourier transform can be formulated simply as a Conv-1d operation with a fixed specific kernel function, and ISTFT as well can be simply implemented with ConvTranspose-1d. We use the fixed kernel function with hamming window (similar to the traditional calculation of STFT). The equation of the STFT kernel function is defined in [19]. Therefore, with implementing traditional STFT and ISTFT with
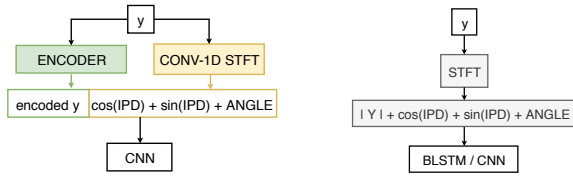
Figure 1: *Multi-channel speech separation; (left) waveform separation, (right) spectrogram separation.*

Conv-1d and ConvTranspose-1d, uPIT-MSE loss can be replaced easily with uPIT-SiSNR. For computing loss function, first need to call ConvTranspose-1d using mix signal phase, and estimate time-samples and then compute Si-SNR.

4. Multi-Channel Speech Separation

As described in subsection 2.2, we simulated multi-channel spatialized WSJ0-2mix for our evaluations. In this section, the integration of multi-channel data into waveform and spectrogram separation networks is described.

In subsection 4.1 we introduce the spatial features, i.e., IPD, used in our multi-channel framework. IPD (inter-microphone phase difference) features are representing the correlation of multi-channel signals and are useful for spatial locating of overlapping sources [3]. Next, in subsection 4.2, angle features are explained which we used them to locate the speakers and extract the speech of a target speaker [3].

4.1. Single-Channel vs Multi-Channel Separation

IPD (inter-microphone phase difference) features represent spatial location information [3] and are calculated as,

$$\text{IPD}_{i,t,f} = \angle(\frac{Y_{i_1,t,f}}{Y_{i_2,t,f}}), i = 1:6 \qquad (3)$$

where i represents an entry in a microphone pair list defined for calculating the IPD; and i_1 and i_2 are the index of microphones in each pair. We calculate IPD features for 6 pairs and then concatenate 6-cos(IPD) and 6-sin(IPD) together.

Figure 1 shows the integration of IPD features (as well as angle features described in the next subsection) into spectrogram and waveform separation networks. For spectrogram separation, multi-channel framework is straightforward; the IPD features are appended with $|Y_0|$, i.e., magnitude of channel 0, and then are fed into the network. For waveform separation, the time-samples at the input are analyzed with Conv-1d implementation of STFT and IPD features are calculated; then they are appended to the encoder output, and followed by the same separation network. However, concatenation of IPD features in time-domain networks can be applied with various fusion methods, which is explored more in detail in [19].

4.2. Target Speaker Extraction vs uPIT Separation

Multi-channel separation networks can also use speakers location information to extract the target speaker of interest. Here, we use angle features [3] for this purpose.

Speech separation networks need to have a prior knowledge of number of overlapping speakers. Therefore, for different conditions when number of overlapping speakers are changing or is unknown, we need to train different models (as these systems usually estimate one mask function for each speaker simultaneously). Using directional information or speaker-related knowledge can help to extract a target speaker; and with a single trained model and multiple passes can extract all overlapping sources. Therefore, we extend our multi-channel frame-

work with angle features (A) [3] to locate a target speaker and extract his/her speech; the block-diagram is shown in Figure 1. These directional features are formulated as,

$$A_{s,t,f} = \sum_{i=1}^{6} \frac{e_s^{i,f} \frac{Y_{i_1,t,f}}{Y_{i_2,t,f}}}{|e_s^{i,f} \frac{Y_{i_1,t,f}}{Y_{i_2,t,f}}|}, \qquad (4)$$

where s is the speaker index and $e_s^{i,f}$ represents steering vector coefficient of speaker s direction of arrival at microphone i for frequency f [3]. In our experiments, we append both speaker angle features together, and always assume the first one is the target speaker for optimizing the target extraction loss.

For traditional uPIT loss function, where S speakers are separating simultaneously, their reference outputs are required for optimizing the network. However, for TGT (target) loss only the reference speech of the target speaker is required. Using more informative directional and speaker related features are studied further in [20].

5. Experiments

5.1. Experimental Setup

Table 1 summarizes the network setup and details of models we applied in our system development (F- and T- mean frequency and time-domain in the tables). Parameters L (length of filters in encoder), N (number of filters in encoder), X (number of convolutional blocks in each repeat), R (number of repeats), P (kernel size in convolutional blocks), B (number of channels in bottleneck 1×1-conv block) and H (Number of channels in convolutional blocks) have the same definition as used in [16]. Please refer to [16] for more details on the structure of the separation network and its hyper-parameters. For all of our time-domain networks hyper-parameters X/R/B/H/P=8/4/257/512/3 are fixed, and for the rest details are included in the Table 1. The gLN and BN in Table 1 are also global layer normalization and batch normalization. In addition, the selected pairs for IPDs are (1, 4), (2, 5), (3, 6), (1, 2), (3,4) and (5, 6) in all experiments.

We used KALDI [21] ASR recipe for WSJ (trained on clean non-reverberate data) to calculate the WERs. For reference, the mix signal WER is 89.91%, and we report the WER reduction rate obtained with each developed speech separation system. In addition, scores for oracle ideal binary mask (IBM), ideal amplitude mask (IAM), ideal ratio mask (IRM) and ideal phase sensitive mask (IPSM) are all reported.

5.2. Experimental Result

We evaluate our proposed solutions in three different evaluation sets. First, incorporating CNN separation network as well as Si-SNR loss function in spectrogram separation is evaluated. Table 2 summarizes the results. This table only includes results from single-channel models. Results confirm the effectiveness of the Si-SNR loss function over MSE. However, for the single-channel input, BLSTM structure slightly performs better than CNN structure.

Next, for time-frequency domain masking network we employ IPD and angle features to extract a target speaker. We evaluate the effectiveness of the multi-channel framework with Si-SNR loss function, comparing BLSTM and CNN separation network structures here. Table 3 shows the results. Here, uPIT-SiSNR and TGT-SiSNR are compared against each other. The results confirm the effectiveness of the proposed multi-channel framework. However, better directional features [20] may be applied here to gain better performance on target speaker extraction.

Finally, we compare waveform separation against spec-

Table 1: *Experimental setup for time/frequency-domain models.*

Model	Input	# of Parameters	Setting	Normalization
F-CNN-1	$\|Y_0\|$	8.78M	N=257	gLN
F-CNN-2	$\|Y_0\|$ + cos(IPD) + sin(IPD)	9.58M	N=257×13	gLN
F-CNN-3	$\|Y_0\|$ + cos(IPD) + sin(IPD) + Angle	9.71M	N=257×15	gLN
F-BLSTM-1	$\|Y_0\|$	67.05M	4 × BLSTM-896	-
F-BLSTM-2	$\|Y_0\|$ + cos(IPD) + sin(IPD)	89.15M	4 × BLSTM-896	-
F-BLSTM-3	$\|Y_0\|$ + cos(IPD) + sin(IPD) + Angle	92.84M	4 × BLSTM-896	-
T-CNN-1	y_0	8.76M	L/N=40/256	gLN
T-CNN-2	y_0 + cos(IPD) + sin(IPD)	8.83M	L/N=40/256	BN

Table 2: *Comparing CNN separation network and Si-SNR loss function in spectrogram separation with single-channel input.*

Model	Loss	Si-SNR					SDR				
		0-15°	15-45°	45-90°	90-180°	AVG	0-15°	15-45°	45-90°	90-180°	AVG
F-BLSTM-1	uPIT-SiSNR	7.54	7.80	7.72	7.81	7.74	8.14	8.39	8.29	8.38	8.32
	uPIT-MSE	6.50	6.79	6.67	6.78	6.71	6.98	7.24	7.11	7.22	7.16
F-CNN-1	uPIT-SiSNR	7.08	7.48	7.45	7.48	7.42	7.70	8.06	8.02	8.06	8
	uPIT-MSE	6.31	6.67	6.54	6.65	6.58	6.81	7.13	7.0	7.1	7.04

Table 3: $|Y_0|$+cos(IPD)+sin(IPD)+Angle *are input to the network to evaluate the performance of multi-channel framework in extracting the speaker of interest.*

Model	Loss	Si-SNR					SDR				
		0-15°	15-45°	45-90°	90-180°	AVG	0-15°	15-45°	45-90°	90-180°	AVG
F-BLSTM-3	uPIT-SiSNR	5.09	9.11	9.62	9.31	8.72	5.81	9.64	10.11	9.85	9.27
	TGT-SiSNR	4.59	9.07	9.59	9.13	8.58	5.31	9.60	10.08	9.64	9.12
F-CNN-3	uPIT-SiSNR	5.59	9.38	10.29	10.79	9.49	6.92	10.36	11.2	11.75	10.50
	TGT-SiSNR	4.88	9.69	10.32	9.80	9.2	5.61	10.19	10.77	10.27	9.71

Table 4: *Comparing spectrogram and waveform separation for both separation and ASR tasks.*

# Channels	Domain	Model	Si-SNR					SDR					PESQ	WER Reduc. (%)
			0-15°	15-45°	45-90°	90-180°	AVG	0-15°	15-45°	45-90°	90-180°	AVG		
1-ch	time	T-BLSTM	-	-	-	-	-	-	-	-	-	-	-	-
		T-CNN-1	9.02	9.33	9.59	9.71	**9.47**	9.57	9.83	10.09	10.2	**9.97**	1.95	45.53
	freq	F-BLSTM-1	7.54	7.80	7.72	7.81	**7.74**	8.14	8.39	8.29	8.38	**8.32**	1.77	32.21
		F-CNN-1	7.08	7.48	7.45	7.48	**7.42**	7.70	8.06	8.02	8.06	**8**	1.77	35.17
m-ch	time	T-BLSTM	-	-	-	-	-	-	-	-	-	-	-	-
		T-CNN-2	7.70	11.63	12.33	12.62	**11.55**	8.31	12.07	12.74	13.03	**11.99**	2.10	59.11
	freq	F-BLSTM-2	5.41	9.37	10.13	10.65	**9.38**	6.13	9.89	10.62	11.13	**9.91**	1.92	45.32
		F-CNN-2	6.88	10.27	11.02	11.54	**10.36**	7.5	10.75	11.47	11.99	**10.84**	2.00	51.73
		Oracle IBM	11.56	11.51	11.53	11.53	**11.53**	11.93	11.86	11.88	11.88	**11.88**	2.01	50.37
		Oracle IAM	11.05	11.03	11.05	11.03	**11.04**	11.33	11.3	11.31	11.29	**11.30**	2.23	71.45
		Oracle IRM	11.01	10.96	10.98	10.97	**10.98**	11.45	11.39	11.39	11.39	**11.40**	2.22	70.28
		Oracle IPSM	13.68	13.6	13.64	13.63	**13.63**	14.04	13.94	13.98	13.97	**13.98**	2.28	71.10
		Reference											2.35	73.01

trogram separation, for both single-channel and multi-channel frameworks. Table 4 (in the table 1-ch means single channel input, m-ch means multi-channel input) shows the results including all Si-SNR, SDR, PESQ and WER scores to not only compare their performance for speech separation task, but also evaluate the impact of them on a speech recognition task. All models in this table are trained with uPIT-SiSNR loss function. The results confirm that for multi-channel (using $|Y_0|$ + cos(IPD) + sin(IPD)) models, CNN separation structure helps to decrease the gap further between time and frequency domain models. In all tables, Si-SNR and SDR scores are reported on 4 sets of angle-differences between the two overlapping speakers as well; which results show that for smaller angle differences, more informative inputs are required.

5.3. Discussion

Based on the results summarized in Tables 2, 3, 4, we can confirm the effectiveness of CNN-based separation structure in frequency-domain. Although, for single-channel inputs F-BLSTM-1 slightly outperforms F-CNN-1 in terms of SDR and Si-SNR; in terms of WER, F-CNN-1 performs better and for multi-channel framework F-CNN-2 achieves better results than F-BLSTM-2 for all criteria. Moreover, Multi-channel models always outperform single-channel models in terms of all reported criteria; F-BLSTM-2, F-CNN-2, and T-CNN-2 relatively gain +19.11%, +35.5%, and +20.26% improvement in terms of SDR criterion, respectively. F-CNN-2 decreases the gap between frequency-domain (with using phase of the mix signal) and time-domain separators; specially in terms of PESQ, results are competitive, 2.1 for T-CNN-2 and 2 for F-CNN-2. In

sum, utilizing multi-channel framework can benefit speech separation task. Although, frequency-domain separation improved with Si-SNR and better separation network design; it can improve further with using better phase component at the reconstruction layer.

6. Conclusions

In this study, multi-channel spatialized reverberant WSJ0-2mix dataset prepared and studied for alternative speech separation setups; including single and multi-channel models, target speaker extraction framework, different loss functions for both spectrogram and waveform input representations. We improved the performance of frequency-domain speech separation with incorporating Si-SNR loss function (directly optimizing the separation criterion). In addition, we explored replacement of BLSTM separation network with CNN structure, which improved the performance for multi-channel inputs, with almost 10 times less number of parameters. We reported PESQ, WER, SDR and Si-SNR scores, to not only evaluate the performance of the introduced models on separation task, but also on speech recognition task as well. Incorporating separation network and loss function from TasNet into frequency-domain shown to be effective; with using CNN-separator, SDR relatively improved by up to +9% on multi-channel scenario; and with Si-SNR loss function SDR improved by up to +16% for single-channel models. With multi-channel framework, SDR, Si-SNR, PESQ and WER relatively improved over the single-channel framework by up to +35.5%, +39.62%, +12.99% and +46%, respectively.

7. References

[1] 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.

[2] S. Haykin and Z. Chen, "The cocktail party problem," *Neural computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[3] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 558–565.

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

[5] M. Stark, M. Wohlmayr, and F. Pernkopf, "Source–filter-based single-channel speech separation using pitch information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 242–255, 2011.

[6] T. Kristjansson, J. Hershey, P. Olsen, S. Rennie, and R. Gopinath, "Super-human multi-talker speech recognition: The ibm 2006 speech separation challenge system," in *Ninth International Conference on Spoken Language Processing*, 2006.

[7] 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.

[8] P. Divenyi, *Speech separation by humans and machines*. Springer Science & Business Media, 2004.

[9] K. Hu and D. Wang, "An unsupervised approach to cochannel speech separation," *IEEE Transactions on audio, speech, and language processing*, vol. 21, no. 1, pp. 122–131, 2013.

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

[11] 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.

[12] 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 *IEEE ICASSP*, 2018, pp. 6–10.

[13] C. Weng, D. Yu, M. L. Seltzer, and J. Droppo, "Deep neural networks for single-channel multi-talker speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 10, pp. 1670–1679, 2015.

[14] Z.-Q. Wang, J. L. Roux, D. Wang, and J. R. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," *arXiv preprint arXiv:1804.10204*, 2018.

[15] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5739–5743.

[16] Y. Luo and N. Mesgarani, "Tasnet: Surpassing ideal time-frequency masking for speech separation," *arXiv preprint arXiv:1809.07454*, 2018.

[17] 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.

[18] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[19] R. Gu, J. Wu, S. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "End-to-end multi-channel speech separation," in *https://arxiv.org/abs/1905.06286*, 2019.

[20] R. Gu, L. Chen, S. Zhang, J. Zheng, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Neural spatial filter: Target speaker separation using directional information," in *ISCA INTERSPEECH*, 2019.

[21] 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," IEEE Signal Processing Society, Tech. Rep., 2011.