IMPROVING SPEAKER DISCRIMINATION OF TARGET SPEECH EXTRACTION WITH TIME-DOMAIN SPEAKERBEAM

Marc Delcroix[1], *Tsubasa Ochiai*[1], *Katerina Zmolikova*[*2], *Keisuke Kinoshita*[1], *Naohiro Tawara*[1], *Tomohiro Nakatani*[1], *Shoko Araki*[1]

[1]NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
[2]Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Czechia

ABSTRACT

Target speech extraction, which extracts a single target source in a mixture given clues about the target speaker, has attracted increasing attention. We have recently proposed SpeakerBeam, which exploits an adaptation utterance of the target speaker to extract his/her voice characteristics that are then used to guide a neural network towards extracting speech of that speaker. SpeakerBeam presents a practical alternative to speech separation as it enables tracking speech of a target speaker across utterances, and achieves promising speech extraction performance. However, it sometimes fails when speakers have similar voice characteristics, such as in same-gender mixtures, because it is difficult to discriminate the target speaker from the interfering speakers. In this paper, we investigate strategies for improving the speaker discrimination capability of SpeakerBeam. First, we propose a time-domain implementation of SpeakerBeam similar to that proposed for a time-domain audio separation network (TasNet), which has achieved state-of-the-art performance for speech separation. Besides, we investigate (1) the use of spatial features to better discriminate speakers when microphone array recordings are available, (2) adding an auxiliary speaker identification loss for helping to learn more discriminative voice characteristics. We show experimentally that these strategies greatly improve speech extraction performance, especially for same-gender mixtures, and outperform TasNet in terms of target speech extraction.

Index Terms— Target speech extraction, time-domain network, spatial features, multi-task loss

1. INTRODUCTION

Recently, deep learning based speech separation approaches have attracted increasing attention [1–4]. Earlier approaches such as deep clustering [1] and permutation invariant training (PIT) [2], performed processing in the frequency-domain and generated time-frequency masks for each source in the mixture. More recently, a convolutional time-domain audio separation network (Conv-TasNet) has been proposed and led to great separation performance improvement surpassing ideal time-frequency masking [4–6]. The separation performance of TasNet has been further improved by exploiting spatial information when a microphone array is available [7].

Despite the great success of neural network-based *speech separation*, it requires knowing or estimating the number of sources in the mixture and still suffers from a global permutation ambiguity issue, i.e. an arbitrary mapping between source speakers and outputs. These limitations arguably limit the practical usage of speech separation. In contrast, *target speech extraction* exploits an auxiliary clue to identify the target speaker in the mixture and extracts only speech of that speaker. After our initial work [8, 9], research on target speech extraction has gained increasing attention [10–14], as it naturally avoids the global permutation ambiguity issue and does not require knowing the number of sources in the mixtures.

We have proposed SpeakerBeam [8, 9], which is a target speech extraction method that exploits a speaker embedding vector derived from an adaptation or enrollment utterance of the target speaker to guide a neural network towards extracting speech of that speaker. This is realized by combining two networks, a sequence summary network [15] that computes the speaker embedding vector from the amplitude spectrum of the adaptation utterance and a speech extraction network that accepts the amplitude spectrum of the speech mixture and the embedding vector as inputs and generates a time-frequency mask for extracting the target speaker. In this paper, we call this approach frequency-domain SpeakerBeam (FD-SpeakerBeam).

We have shown that FD-SpeakerBeam could achieve competitive speech extraction performance and be used as a front-end for automatic speech recognition (ASR) [9]. However, we observe a great performance gap between same-gender and different-gender mixtures [16]. It is indeed difficult to discriminate a target speaker in a mixture when speakers have similar voice characteristics.

In this paper, we investigate strategies to tackle this issue. First, following the success of TasNet, we propose *a time-domain implementation of SpeakerBeam* (TD-SpeakerBeam), whose speech extraction network accepts time-domain signals of the mixture, and outputs directly the time-domain signal of the target speaker. We also replace the sequence summary network with a convolutional network to obtain richer speaker embedding vectors.

Moreover, to further improve speaker discrimination capability, we extend TD-SpeakerBeam to accept spatial information from microphone array recordings as additional input features. We argue that simply adding spatial features to the input of TD-SpeakerBeam may limit the potential to process spatial information. Consequently, we propose an alternative approach, called *internal combination*, for exploiting spatial information more effectively within the SpeakerBeam framework.

Finally, to enforce learning more discriminative speaker embedding vectors, we propose using a multi-task loss for training SpeakerBeam, that combines a speech reconstruction loss with a *speaker identification loss* (SI-loss).

We performed experiments on two datasets, which show that (1) TD-SpeakerBeam greatly improves target speech extraction performance and outperforms a competitive system based on TasNet separation followed by an x-vector [17] based target speech selection module, (2) exploiting spatial features with the proposed internal combination helps target speech extraction especially for same-gender mixtures, (3) the additional SI-loss consistently improves

*Katerina Zmolikova was partially supported by the Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602".

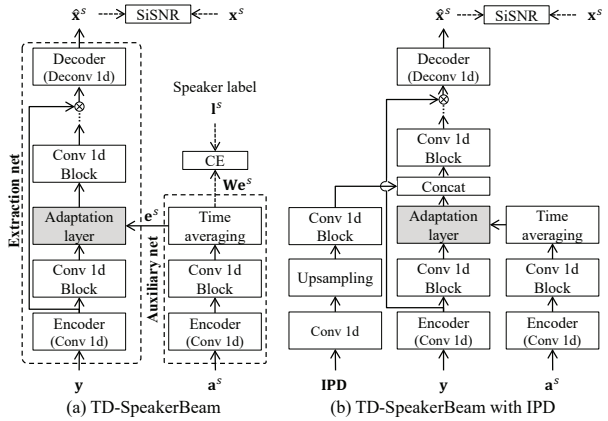


Fig. 1. Block diagram of (a) proposed TD-SpeakerBeam and (b) TD-SpeakerBeam with internal combination of IPD features. The inputs and outputs of the networks are defined in Section 2.1

performance when a sufficient number of speakers are included in the training data, (4) by varying the number of training speakers, although TasNet performance does not change significantly, SpeakerBeam benefits greatly from more training speakers especially for same-gender mixtures because it helps improving target speaker identification. These results confirm the efficiency of the proposed strategies for improving the target speaker discrimination capability of SpeakerBeam.

2. PROPOSED TIME-DOMAIN SPEAKERBEAM

Let us first describe the implementation of TD-SpeakerBeam. Then, in section 2.2, we discuss approaches for exploiting spatial information when microphone array recordings are available. Finally, in section 2.3, we introduce a multi-task loss to improve speaker discrimination even when only a single microphone is available.

2.1. TD-SpeakerBeam

Figure 1-(a) is a block diagram of the proposed TD-SpeakerBeam. Let $\mathbf{y}$, $\mathbf{a}^s$ and $\hat{\mathbf{x}}^s$ be the time-domain signals of the speech mixture, the adaptation utterance, and the estimated target speech for target speaker s. SpeakerBeam is composed of two networks, an *extraction network*, and an *auxiliary network*. In the original FD-SpeakerBeam [9], these networks accept the amplitude spectrum of the mixture and adaptation signals as inputs and generate a time-frequency mask. In this paper, we modify the implementation of these networks to input and output time-domain signals.

The time-domain extraction network follows a similar configuration as Conv-TasNet [5], i.e. it consists of a 1d convolution layer that accepts the mixture signal $\mathbf{y}$ (encoder layer), several convolution blocks, and finally, a 1d deconvolution (decoder layer) that outputs the extracted speech signal in the time-domain, $\hat{\mathbf{x}}^s$.

There are two major differences with Conv-TasNet. First, the output consists of a single signal corresponding to the target speech only. Second, we insert an adaptation layer between the first and second convolution blocks[1] to drive the network towards extracting the target speech. The adaptation layer accepts a speaker embedding vector of the target speaker, $\mathbf{e}^s$, as auxiliary information. We use a multiplicative adaptation layer following our previous work [16], although other adaptation layers could be used [9, 13].

[1]We found in preliminary experiments that placing the adaptation layer after the first convolution block achieved the best performance.

The target speaker embedding vector, $\mathbf{e}^s$, is computed by the time-domain auxiliary network. In the original FD-SpeakerBeam, the auxiliary network consists of a sequence summary network [15], i.e. a few fully connected layers followed by a time-averaging operation. Here, we propose using a convolutional auxiliary network to accept the time-domain input signal of the adaptation utterance $\mathbf{a}^s$. The auxiliary network consists of an encoder layer and a single convolution block similar to those used in the extraction network.

2.2. Spatial features

Spatial information extracted from multi-channel recordings can provide an alternative source of information about the mixtures that could help discriminate speakers better. There have been several works showing the benefit of adding spatial features to the input of speech enhancement networks [7, 18, 19]. For example, [7] recently showed that the inter-microphone phase difference (IPD) features could improve the separation performance of TasNet in reverberant conditions. The IPD of the mixture signal between two microphones is defined as,

$$\Phi_{i,j,t,f} = \angle \left(\frac{Y_{i,t,f}}{Y_{j,t,f}} \right), \tag{1}$$

where $Y_{i,t,f} \in \mathbb{C}$ is the short-time Fourier transform (STFT) coefficient of the mixture signal at microphone i, t is the time-frame index, and f is the frequency index. Here we limit our investigation to the two-microphone case. Following [7], we use cosine and sine of the IPDs as spatial features,

$$\begin{aligned} \mathbf{IPD}_t = &[\cos(\Phi_{1,2,t,1}), \ldots, \cos(\Phi_{1,2,t,F}), \\ &\sin(\Phi_{1,2,t,1}), \ldots, \sin(\Phi_{1,2,t,F})], \end{aligned} \tag{2}$$

where F is the number of frequency bins. Note that the frame size and window shift of the STFT used to compute the IPD features may differ from the window size and shift used in the encoder of the extraction network. IPD features are thus upsampled to match the settings of the extraction network

IPD features provide spatial information related to the direction of sources in the mixture. SpeakerBeam extracts the target speech based on the speaker embedding vector, $\mathbf{e}^s$, that may represent "spectral" information[2] about the target speaker, but does not include spatial information. Consequently, it is not obvious how to efficiently combine the IPD features and the target speaker embedding vector as they represent different information. In this paper, we consider two schemes, *input combination* and *internal combination*.

The *input combination* is similar to that proposed for TasNet in [7], where the IPD features (processed with a convolutional layer and upsampled) are concatenated to the output of the encoder layer of the extraction network. Input combination may force the initial convolution block to combine spatial information from the IPD features and "spectral" information from the mixture signal $\mathbf{y}$ into a "spectral" representation, which allows the adaptation layer (coming after the first convolutional block) to select the target speech by comparing this "spectral" representation with the target speaker embedding vector, $\mathbf{e}^s$. This may reduce the potential of the network to fully exploit spatial information by the upper layers of the network.

Figure 1-(b) shows a schematic diagram of TD-SpeakerBeam with the alternative *internal IPD combination*. It combines the IPD features (processed with a 1D convolutional layer, upsampling, and a convolution block) after the adaptation layer. Therefore, this lets the speaker selection operate based only on the "spectral" information, and the spatial information can be exploited by the upper layers without being obstructed by the adaptation layer.

[2]Strictly speaking it is not the usual spectrum information as we use a learnable convolutional encoder layer to analyze the signal instead of STFT.

Table 1. Amount of data and number of female (F) and male (M) speakers.

	Train				Test			
	#Mixtures	#Spks	#F	#M	#Mixtures	#Spks	#F	#M
WSJ	20k	101	52	49	3k	18	8	10
CSJ	50k	937	166	771	15k	30	10	20

Here, we only consider exploiting spatial information as additional information to the extraction network. Besides, SpeakerBeam can also be combined with beamforming, which is particularly efficient for ASR applications [8, 20], but is out of the scope of this paper.

2.3. multi-task learning with additional SI-loss

The extraction network and auxiliary networks are trained jointly from random initialization given the speech mixtures, $\mathbf{y}$, adaptation utterances, $\mathbf{a}^s$, and the target speech signals $\mathbf{x}^s$. In our previous works [8, 9], we trained SpeakerBeam using only a target speech reconstruction loss. In this paper, we propose using a multi-task loss for training TD-SpeakerBeam that combines scale-invariant source-to-noise ratio (SiSNR) [21] as signal reconstruction loss and cross-entropy-based SI-loss. The SI-loss is used to obtain more discriminative speaker embedding vectors. The multi-task loss is given by,

$$L(\Theta|\mathbf{y}, \mathbf{a}^s, \mathbf{x}^s, \mathbf{l}^s) = -\text{SiSNR}(\mathbf{x}^s, \hat{\mathbf{x}}^s) + \alpha \text{CE}(\mathbf{l}^s, \sigma(\mathbf{W}\mathbf{e}^s)), \quad (3)$$

where Θ are the model parameters, and $\mathbf{l}^s$ is a one-hot vector representing the target speaker ID, $\text{SiSNR}(\mathbf{x}^s, \hat{\mathbf{x}}^s)$ is the SiSNR between the estimated and true target speech, $\text{CE}(\mathbf{l}^s, \sigma(\mathbf{W}\mathbf{e}^s))$ is the cross entropy between the speaker label $\mathbf{l}^s$ and the speaker embedding vector projected onto the training speaker space, $\mathbf{W}\mathbf{e}^s$, $\mathbf{W}$ is a projection matrix, $\sigma(\cdot)$ is a softmax function, and α is a scaling parameter.

3. RELATED PRIOR WORK

An alternative way to perform target speech extraction consists of performing speech separation followed by target speaker selection from the separated signals. Such a scheme was proposed in [22] for deep attractor network [23], but to the best of our knowledge has not been investigated for time-domain separation approaches. In the experiments, we compare TD-SpeakerBeam with TasNet separation followed by x-vector-based speaker selection [17], which can be considered a strong baseline for target speech extraction.

We borrowed from previous works on multi-channel source separation [7, 19] that IPD features may be good candidates for increasing extraction performance. Besides adding IPD features to the extraction network, an alternative approach was recently proposed [14], where a set of fixed beamformers combined with an attention module on the output of the beamformers was used to perform a rough initial target speech extraction followed by a refinement step with FD-SpeakerBeam. Investigating such a scheme with TD-SpeakerBeam or other spatial features will be part of our future works.

4. EXPERIMENTS

4.1. Datasets

We performed experiments using two datasets, multi-channel WSJ0 2 mixtures (MC-WSJ0-2 mix) and CSJ-2mix. Table 1 shows details of the amount of data and the number of female and male speakers in the training and test sets.

MC-WSJ0-2 mix is a publicly available multi-channel version of the WSJ0-2mix corpus [24] that consists of mixtures of WSJ0 utterances [25]. Multi-channel recordings are generated by convolving clean speech signals with room impulse responses simulated with the image method for reverberation time of up to about 600ms. The dataset consists of 8 channel recordings, but we use only 2 channels in our experiments. This dataset has only 101 training speakers. We use it thus only for the investigations on the use of spatial features.

The second dataset consists of single-channel 2-speaker mixtures that we simulated by mixing utterances from the corpus of spontaneous Japanese (CSJ) [26] at SNR between -5 and 5 dBs. This dataset has a larger number of training speakers (937 speakers) and is used to investigate the effect of the SI-loss and the impact of the number of training speakers.

For both datasets, we randomly selected adaptation utterances of the target speaker in a mixture from the utterances of that speaker that differed from the utterance in the mixture. In the MC-WSJ0-2mix experiments, the adaptation utterances did not contain reverberation, although a similar level of performance could be achieved with reverberant adaptation utterances. We used an 8kHz sampling frequency for all our experiments.

4.2. Experimental settings

TD-SpeakerBeam was implemented based on the open source Conv-TasNet implementation [27]. In particular, following the hyper-parameter notations in the original paper [5], we set the hyper-parameters to N=256, L=20, B=256, H=512, P=3, X=8, R=4. The auxiliary network consisted of an encoder layer and a single convolution block.

We compare the proposed TD-SpeakerBeam with (1) TasNet with oracle target speech selection, (2) TasNet with x-vector-based target speech selection, and (3) our previous implementation of FD-SpeakerBeam. TasNet used the network configuration described in [5, 27], with hyper-parameters equivalent to those of TD-SpeakerBeam. Oracle speaker selection was performed by finding the speaker permutation that maximizes the signal-to-distortion ratio (SDR). For TasNet with x-vector-based speaker selection [17], we used the same TasNet network but selected the target speech as the output of the TasNet separation module whose x-vector presented the highest cosine similarity with the x-vector of the adaptation utterance. We used the publicly available x-vector extractor model that was trained on multi-condition noisy and reverberant data to extract x-vectors [28, 29].

The network architecture of FD-SpeakerBeam consisted of 3 BLSTM layers followed by a sigmoid layer and 3 fully connected layers for the auxiliary network. FD-SpeakerBeam was trained with the MSE loss between the amplitude spectrum of clean target speech and masked signals. Details of the configuration can be found in [9]. Note that many aspects of the network configuration and the training procedure differ from that of TD-SpeakerBeam. Consequently, the results of FD-SpeakerBeam are only indicative of the level of performance achieved in our previous works. A more fair comparison between the impact of working in the time and frequency domain in the context of speech separation can be found in [7].

For experiments with MC-WSJ0-2mix, we extracted IPD features using an STFT window of 32 msec and a shift of 16 msec. We compare TasNet with input IPD combination [7] and TD-SpeakerBeam with input and internal IPD combination.

All time-domain models were trained to optimize the SiSNR criterion only (i.e. $\alpha = 0$ in Eq. (3)) except when we mentioned the use of the SI-loss, in which case we used $\alpha = 10$. As the evaluation metrics, we used the scale-invariant SDR of BSSeval [30]

Table 2. SDR (dB) on the MC-WSJ0-2mix corpus. Bold-fonts indicate best performance (except for oracle).

	IPD	FF	MM	FM	avg
(1) Mixture	-	0.17	0.16	0.16	0.16
(2) TasNet (oracle)	-	*8.68*	*9.75*	*12.14*	*10.84*
(3)	input	*11.52*	*11.37*	*12.17*	*11.83*
(4) TasNet (xvect)	-	4.59	4.93	11.44	8.35
(5)	input	6.01	5.80	11.35	8.80
(6) FD-SpkBeam	-	5.19	5.32	10.27	7.94
(7) TD-SpkBeam	-	9.13	9.47	**12.77**	11.17
(8)	input	9.02	9.71	12.55	11.11
(9)	internal	**10.17**	**10.30**	12.49	**11.45**

Table 3. SDR [dB] on the CSJ-2mix corpus.

	FF	MM	FM	avg
(1) Mixture	0.19	0.18	0.18	0.18
(2) TasNet (oracle)	*11.86*	*14.81*	*17.01*	*15.28*
(3) TasNet (xvect)	7.65	12.51	16.29	13.35
(4) FD-SpkBeam (Freq)	6.42	8.35	10.52	8.93
(5) TD-SpkBeam	12.56	17.15	18.83	17.24
(6) TD-SpkBeam + SI-loss	**13.60**	**17.75**	**19.22**	**17.81**

4.3. Results with IPD features using MC-WSJ0-2mix

Table 2 shows the SDR for the MC-WSJ0-2mix experiments for mixtures of female-female (FF), male-male (MM) and female-male (FM) speakers. We confirmed that TasNet with oracle target speaker selection (row (2)) achieved high SDR performance. Moreover, TasNet with input combination of IPD features (row (3)) further improved performance especially for mixtures of same-gender speakers. These results can be considered an upper-bound for TasNet-based target speaker extraction. We omitted results with the internal IPD combination for TasNet, as it performed worse than using IPD features at the input.

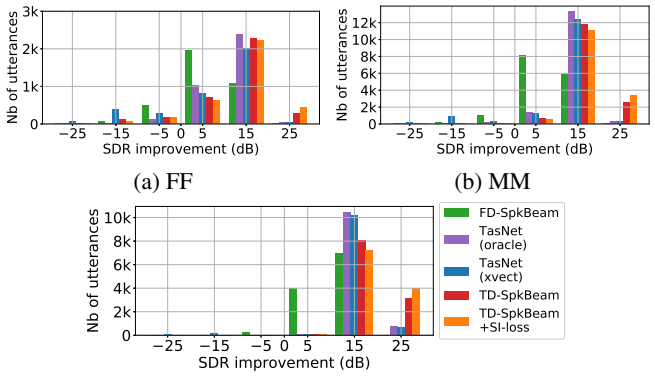
Performance dropped greatly when using x-vector-based speaker selection (row (4) and (5)), especially for FF and MM cases. Although the x-vector extractor was trained on multi-condition data, it may still be affected by reverberation, which may partly contribute to the poor performance. However, reverberation is not the only reason for the performance drop because x-vector selection performed significantly worse than oracle even for the following CSJ experiments that do not include reverberation.

FD-SpeakerBeam (row (6)) performed slightly worse than TasNet (xvect). The proposed TD-SpeakerBeam (row (7)) outperformed all systems but TasNet with oracle speaker selection. Especially, there is a smaller performance gap between mixtures of speakers of the same and different genders than with FD-SpeakerBeam. We further confirmed that TD-SpeakerBeam with internal IPD combination (row (9)) improved performance by up to 1 dB and performed better than input combination (row (8)).

4.4. Results with the SI-loss on CSJ-2mix

Table 3 shows the SDR for TasNet with oracle and x-vector-based speaker selection, FD-SpeakerBeam and TD-SpeakerBeam without and with SI-loss. These results were obtained when using all 937 training speakers for training the models. TD-SpeakerBeam achieved much better performance than FD-SpeakerBeam and TasNet with or without oracle speaker selection. Moreover, SI-loss provided further consistent performance improvement of up to 1 dB.

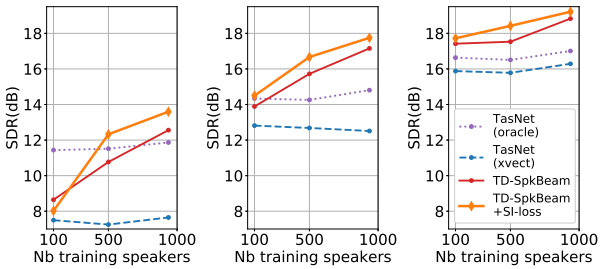
Figure 2 shows the histogram of the SDR improvement for FF, MM and FM mixtures. TD-SpeakerBeam with or without SI-loss greatly reduced processing failures (SDR improvement of 0 dB or less). Moreover, the SI-loss led to better overall performance (more results with high SDR improvement).



(a) FF (b) MM

(c) FM

Fig. 2. Histogram of the SDR improvement for CSJ 2 mix experiments with 937 training speakers.



(a) FF (b) MM (c) FM

Fig. 3. SDR as a function of the number of training speakers.

Figure 3 shows the SDR as a function of the number of training speakers. The curves were obtained by creating 3 different training sets with 100, 500 and all 937 training speakers. In all cases, we used 50k mixtures. Interestingly, we observe that increasing the number of speakers has little effect on TasNet performance, but greatly improves the performance of SpeakerBeam. This suggests that, for SpeakerBeam, separating signals is somewhat easier than identifying speakers. The SI-loss provided consistent improvement when using more than 100 speakers (This is why we did not use the SI-loss in the MC-WSJ0 experiments). Note that performance remains significantly lower for FF mixtures, partly because there are fewer female speakers in the training set (see table 1) and also because it appears to be more challenging to separate FF mixtures, as also suggested by the lower performance of TasNet in this case.

5. CONCLUSION

In this paper, we proposed different strategies for improving the target speech discrimination capability of SpeakerBeam. We showed that a time-domain implementation greatly improved performance. Moreover, the performance gap between same-gender and different-gender mixtures could be reduced further by exploiting spatial information, using an additional SI-loss, or by increasing the number of training speakers.

In future work, we would like to combine these techniques to tackle more challenging noisy and reverberant mixtures, e.g. [16]. Moreover, we will also investigate other approaches to integrate spatial information [7, 14] and more discriminative SI-losses [31].

6. REFERENCES

[1] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. of ICASSP'16*, 2016, pp. 31–35.

[2] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. ASLP*, vol. 25, no. 10, pp. 1901–1913, 2017.

[3] K. Kinoshita, L. Drude, M. Delcroix, and T. Nakatani, "Listening to each speaker one by one with recurrent selective hearing networks," in *Proc. of ICASSP'18*, 2018, pp. 5064–5068.

[4] Y. Luo and N. Mesgarani, "TasNet: Surpassing ideal time-frequency masking for speech separation," in *Proc. of ICASSP'18*, 2018.

[5] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. ASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.

[6] Z. Shi, H. Lin, L. Liu, R. Liu, S. Hayakawa, and J. Han, "Furcax: End-to-end monaural speech separation based on deep gated (de) convolutional neural networks with adversarial example training," in *Proc. of ICASSP'19*, 2019, pp. 6985–6989.

[7] F. Bahmaninezhad, J. Wu, R. Gu, S.-X. Zhang, Y. Xu, M. Yu, and D. Yu, "A comprehensive study of speech separation: spectrogram vs waveform separation," in *Proc. of Interspeech'19*, 2019, pp. 4574–4578.

[8] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," in *Proc. of Interspeech'17*, 2017, pp. 2655–2659.

[9] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Cernocky, "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[10] 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," in *Proc. of Interspeech'18*, 2018, pp. 307–311.

[11] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. on Graphics*, vol. 37, no. 4, pp. 112:1–112:11, 2018.

[12] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. of Interspeech'19*, 2019, pp. 2728–2732.

[13] C. Xu, W. Rao, E. S. Chng, and H. Li, "Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss," in *Proc. of ICASSP'19*, 2019, pp. 6990–6994.

[14] G. Li, S. Liang, S. Nie, W. Liu, M. Yu, L. Chen, S. Peng, and C. Li, "Direction-aware speaker beam for multi-channel speaker extraction," in *Proc. of Interpseech'19*, 2019.

[15] K. Vesely, S. Watanabe, K. Zmolikova, M. Karafiat, L. Burget, and J. H. Cernocky, "Sequence summarizing neural network for speaker adaptation," in *Proc. of ICASSP'16*, 2016, pp. 5315–5319.

[16] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, S. Araki, and T. Nakatani, "Compact network for SpeakerBeam target speaker extraction," in *Proc. of ICASSP'19*, 2019, pp. 6965–6969.

[17] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Proc. of SLT'16*, 2016, pp. 165–170.

[18] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *Proc. of ICASSP'15*, 2015, pp. 116–120.

[19] Z. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. of ICASSP'18*, 2018.

[20] T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, and S. Araki, "Beam-TasNet: Time-domain audio separation network meets frequency-domain beamformer," in *Proc. of ICASSP'20 (Submitted)*, 2020.

[21] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR-half-baked or well done?," in *Proc. of ICASSP'19*, 2019, pp. 626–630.

[22] L. Drude, T. von Neumann, and R. Haeb-Umbach, "Deep attractor networks for speaker re-identification and blind source separation," in *Proc. of ICASSP'18*, 2018, pp. 11–15.

[23] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. of ICASSP'18*, 2017, pp. 246–250.

[24] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. of ICASSP'18*, 2018, pp. 1–5.

[25] J. Garofolo, "CSR-I (WSJ0) Complete LDC93S6A," `https://catalog.ldc.upenn.edu/ldc93s6a`, 1993.

[26] K. Maekawa, H. Koiso, S. Furui, and I. H., "Spontaneous speech corpus of Japanese," in *Proc. of LREC'00*, 2000, pp. 947–952.

[27] "`https://github.com/funcwj/conv-tasnet`," .

[28] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, and S. Khudanpur, "Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge," in *Proc. of Interspeech'18*, 2018, pp. 2808–2812.

[29] "`https://github.com/iiscleap/DIHARD-2019-baseline`," .

[30] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[31] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," in *Proc. of Interpseech'19*, 2019.