SPEAKER AND DIRECTION INFERRED DUAL-CHANNEL SPEECH SEPARATION

Chenxing Li[1,2], *Jiaming Xu*[1,2†], *Nima Mesgarani*[3], *Bo Xu*[1,2†]

[1]Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Columbia University, New York, NY, USA

ABSTRACT

Most speech separation methods, trying to separate all channel sources simultaneously, are still far from having enough generalization capabilities for real scenarios where the number of input sounds is usually uncertain and even dynamic. In this work, we employ ideas from auditory attention with two ears and propose a speaker and direction inferred speech separation network (dubbed SDNet) to solve the cocktail party problem. Specifically, our SDNet first parses out the respective perceptual representations with their speaker and direction characteristics from the mixture of the scene in a sequential manner. Then, the perceptual representations are utilized to attend to each corresponding speech. Our model generates more precise perceptual representations with the help of spatial features and successfully deals with the problem of the unknown number of sources and the selection of outputs. The experiments on standard fully-overlapped speech separation benchmarks, WSJ0-2mix, WSJ0-3mix, and WSJ0-2&3mix, show the effectiveness, and our method achieves SDR improvements of 25.31 dB, 17.26 dB, and 21.56 dB under anechoic settings. Our codes will be released at https://github.com/aispeech-lab/SDNet.

Index Terms— dual-channel speech separation, speaker and direction-inferred separation, cocktail party problem.

1. INTRODUCTION

In many environments, the auditory scene is composed of several concurrent speech streams with their spectral features overlapping both in space and time. Human auditory system exhibits a remarkable ability to parse these complex scenes. However, background noise, overlapping speech, and reverberation damage the quality and degrade the performance of speech recognition.

Recently, some researchers attempt to alleviate the problem and pay extensive attention to neural network-based speech separation. In the single-channel-based separation task, many methods have achieved state-of-the-art (SOTA) performance, such as frequency domain-based DPCL [1], DANet [2], PIT [3], Chimera++ [4], CBLDNN-GAT [5], SPNet [6], Deep CASA [7] and time domain-based TasNet [8], FurcaPa [9], DPRNN [10]. These methods design the model structure from different perspectives and follow different training strategies, where the factors affecting performances are investigated in depth. However, these methods meet several challenges: an unknown number of sources in the mixture, permutation problem, and selection from multiple outputs.

In order to deal with the situation that the number of sources in mixed speech is unknown, paper [11] incorporates DPCL into the masking-based beamforming and performs separation. OR-PIT [12]

separates only one speaker from a mixture at a time, and the residual signal is sent to the separation model for the recursion to separate the next speaker. An iteration termination criterion is proposed to identify the number of speakers accurately. A speaker-inferred model [13] uses the Seq2Seq-based method [14, 15] to infer speakers. Speaker information is also appended to the output. Auxiliary autoencoding PIT [16] is proposed to further improve the performance across various numbers of speakers.

Speaker-aware-based networks [17–21] try to deal with the problem of permutation and selection from outputs. These methods are interested in recovering a single target speaker while reducing noise and the effect of interfering speakers. The reference speech from the target speaker should be given in advance.

In addition to the single-channel-based methods, multi-channel-based methods can extract additional direction features to further improve the performance, and some methods are proposed to solve the problems of permutation and output selection. Similar to the speaker-aware-based networks, Li et al. [22] use fixed beamformers to transfer the multi-channel mixture into single-channel signals. An attention network is designed to identify the direction of the target speaker and combine the beamformed signals. SpeakerBeam [18] is then applied to separating the enhanced signal. The direction-aware-based method [23] focuses on the target source in a specific direction by using a time domain-based network.

PIT-based methods [3,5,6,8–10] need prior knowledge about the number of speakers and meet the permutation problem. These methods have some shortcomings in real environments. In the existing methods account for identifying the number of outputs, [12] requires iterative operations, which increases system complexity. [11,12] still can not solve the problem of the selection of outputs. Speaker-aware-based methods need to know the target speaker in advance. The speech of other speakers cannot be separated. Besides, in single-channel-based methods, speakers with similar pitch are difficult to be separated. By extending it to multi-channel-based methods, an additional direction feature can be acquired by the network.

In real environments, a source signal has a unique speaker and direction information. We propose a speaker and direction-inferred dual-channel speech separation network (SDNet), which can infer speaker and direction information first and use them as cues to separate speech. Our contributions are listed as follows: (1) We expand single-channel to multi-channel time domain-based separation based on [13]. Spectral and spatial features are fully utilized. (2) Instead of manually extracting channel differences in [25, 26], the channel differences are extracted by the network and can be optimized end-to-end. (3) This network can simultaneously infer speaker and direction information, and the information is fused as a source mask for separation. By dynamically estimating the number of source masks, the network can cope with the problem of the unknown number of outputs. (4) After separation, speaker and direction information are
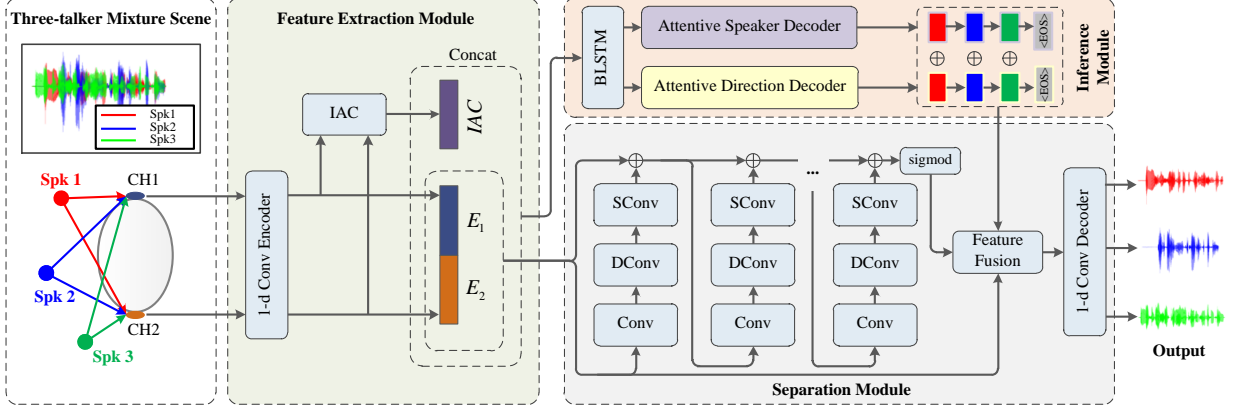

Fig. 1: The model architecture of SDNet. In the separation module, SConv and DConv represent the depth-wise separable convolution [24].

appended to the separated speech. This information can be used in subsequent tasks. The network can deal with the problem of the selection of outputs.

Scale-invariant signal-to-noise ratio (SISNR) [8] and signal-to-distortion ratio (SDR) [27] improvements are used to evaluate the performance. Experimental results show that SDNet can effectively perform separation both on the anechoic and reverberant settings.

2. SYSTEM OVERVIEW

The illustration of our model is shown in Fig. 1. Our network is composed of three components: (1) The feature extraction module processes features from each channel and extracts the channel differences; (2) The inference module parses out the speakers and directions from the mixture and generates source masks; (3) The separation network processes features and integrates the source masks to generate the separated outputs.

2.1. Feature extraction module

2.1.1. convolutional encoder

The encoder transforms the mixture waveforms into an intermediate feature space. In detail, the input segment is transformed into the representation by using a one-dimensional (1D) convolutional layer.

$$E_i = Conv1d(CH_i), \quad i = 1, 2, \tag{1}$$

where E indicates the encoder, and CH_i represents the waveform of i-th channel.

2.1.2. inter-channel attention correlation

Compared with the single-channel-based models, dual-channel-based models can use both spatial and spectral information. This is conducive to the improvement of performance. The time difference between the channels can be obtained by end-to-end training or manually-designed features [25, 26]. We directly calculate the correlation among the channels and integrate it into the network as an additional feature. In detail, similar to the self-attention [28], we calculate the attention correlation between channels. By setting channel 1 as the reference, the inter-channel attention correlation (IAC) is as follows:

$$IAC = \text{softmax}(E_1 E_2^T), \tag{2}$$

where E_1 and E_2 represent the output of encoder 1 and encoder 2, respectively. Channel differences may contribute to the inference module. The feature extraction module has two different outputs, and the outputs are:

$$F = [E_1, E_2], \quad F_o = [IAC, E_1, E_2], \tag{3}$$

where F_o is sent to the inference module, and F is fed into the separation module.

2.2. Inference module

In the inference module, the Seq2Seq-based mechanism [14, 15] is applied to inferring the speakers and directions in a sequence manner. First, the features are mapped into high-level vectors by using stacked bi-directional long short term memory (BLSTM) layers. The specific equation is:

$$h = BLSTM(F_o), \tag{4}$$

where h is the hidden state, and F_o represents the input feature of the inference module.

We use two independent decoding networks to infer the speakers and the directions, respectively. Considering not all speech features make contributions to infer the speakers and directions equally at each step, the attention mechanisms are utilized to produce context vectors by focusing on different portions of the sequence and aggregating the hidden representations. Two attentive decoding networks have similar procedures. Here we formulate the speaker-inferred decoder as follows:

$$\alpha_{ti} = \text{softmax}(tanh(W_1 s_{t-1} + U_1 h_i)), \tag{5}$$

$$c_t = \sum_{i=1}^{T} \alpha_{ti} h_i, \tag{6}$$

where W_1, U_1 are weights, and s_{t-1} is the hidden state of the decoder at time-step $t-1$. c_t is the context vector at time-step t.

For the decoding networks, a global embedding strategy is introduced to alleviate the problem of exposure bias [15], and the embedding feature at t is calculated as follows:

$$e_t^a = \sum_{j=1}^{N} y_{t-1}^j e_j, \tag{7}$$

$$g = \text{sigmoid}(W_2 e_t + U_2 e_t^a), \tag{8}$$

$$e_{s_t} = g \odot e_t + (1 - g) \odot e_t^a, \tag{9}$$

where N is the number of speakers. y_{t-1}^j is the j-th element of y_{t-1} and e_j is the embedding vector of the j-th speaker. e_t^a denotes the weighted average embedding at time t. W_2 and U_2 are weight matrices. e_{s_t} represents the speaker embedding at time t. $\odot$ denotes element-wise multiplication. The hidden state s_t of the decoder at time-step t is computed as follows:

$$s_t = \text{LSTM}(s_{t-1}, [e_{s_t}; c_t]). \tag{10}$$

The final output is calculated as:

$$y_t = \text{softmax}(W_3 f(W_4 s_t + W_5 c_t)), \tag{11}$$

where W_3, W_4, and W_5 are weights. y_t represents the inferred probability distribution of the inferred speaker at time-step t. In each time step, rather than selecting the final output y_t, the speaker embedding, e_{s_t}, is selected as the speaker mask. When the inferred y_t corresponds to an <EOS> (End-of-Sequence), the decoding process is stopped.

The inference process of direction mask, e_{d_t}, is the same as the equations above. The source mask can be obtained as follows:

$$sm_t = e_{s_t} + e_{d_t}, \tag{12}$$

where sm_t means the t-th source mask inferred by the inference module. In the inference module, these two attentive decoders run simultaneously. If one decoder infers an <EOS>, the two decoders are stopped. In the test, the beam search algorithm [29] is applied to finding the top-ranked inference.

2.3. Separation module

Temporal convolutional networks (TCN) [30] effectively memorize long-term dependencies. Dilation rate [30] is used to continuously expand the receptive field. The separation module is the same as the separation module in TasNet [8]. In detail, the streamline of the separation module consists of four convolutional blocks. In each block, for expanding receptive fields, dilated convolutional operations are repeat R times with 1,2,4,..., and 2^{R-1} dilation rates. A sigmoid activation then scales the output.

To generate separated outputs, the decoding process is the inverse process of the encoding layer. It decodes the feature representation to speech samples. Specifically, we use 1D transposed convolution to implement the decoding process:

$$Z_i = F \odot TCN_o \odot sm_i, \quad i = 1, ..., n,$$
$$D(Z)_i = TransposedConv(Z_i), \quad i = 1, ..., n, \tag{13}$$

where TCN_o denotes the output of TCN layers. Z_i represents the high-level feature representatives of i-th inferred source. n is the number of source masks infered in this mixture. $D(\cdot)_i$ represents the i-th separated output.

2.4. Loss function

End-to-end training is performed, and three kinds of loss are adopted: raw-waveform-based SiSNR separation loss, cross-entropy-based speaker-inferred loss, and cross-entropy-based direction-inferred loss. The detailed loss function is formulated as:

$$\mathcal{L} = -\mathcal{L}_{SiSNR-SS} + \lambda \times (\mathcal{L}_{CE-Spk} + \mathcal{L}_{CE-Dir}), \tag{14}$$

Table 1: *The effect of SNet-time in single-channel anechoic datasets and comparison of different methods on SDR improvement (dB).*

System	WSJ0-2mix	WSJ0-3mix	WSJ0-2&3mix
SNet-time	12.35	9.87	10.81
SNet [13]	7.52	5.14	7.05
DPCL++ [31]	10.3	7.1	8.8
uPIT-BLSTM [3]	10.0	7.7	8.9
TasNet [8]	15.0	12.8	—
OR-PIT [12]	15.0	12.9	—

where λ is a hyper-parameter. For the inference module, speaker indexes act as the speaker labels, which are 101 in this experiment. 37 directions are chosen as the direction labels, which are distributed from 0 degrees to 180 degrees with a 5-degree interval. The labels of direction are generated during the data simulation. Meanwhile, <BOS> (Begin-of-Sequence) and <EOS> are also added to the speaker and direction label sets. For each sample, <BOS> is placed at the top of the labels, and <EOS> is placed at the end. <BOS> means that the network starts to infer. <EOS> is used for the network to determine the end of decoding.

3. EXPERIMENTS

3.1. Experimental setup

The proposed methods are evaluated on 8k Hz single and dual-channel WSJ0-2mix, WSJ0-3mix, and WSJ0-2&3mix datasets [1]. For both single-channel and stereo datasets, WSJ0-2mix and WSJ0-3mix contain 30 hours of training data, 10 hours of development data, and 5 hours of test data. The mixing signal-to-noise ratio, pairs, dataset partition are exactly coincident with paper [1]. WSJ0-2&3mix is the union of WSJ0-2mix and WSJ0-3mix. Anechoic and reverberant stereo datasets are generated by convolving the clean speech with the room impulse responses [32, 33]. For reverberant datasets, the reverberation time is uniformly sampled from 40 ms to 200 ms. We place 2 microphones at the center of the room. The distance between microphones is 10 cm. Sound sources are randomly placed in the room. The training set and the test set contain 101 and 18 speakers, respectively. The speakers in the test set are different from the speakers in the training set and the development set. During training, the label order in the inference module is sorted in descending order according to speech energy.

In (inChannel, outChannel, kernel, stride)-format, for the encoder in the feature extraction module, 1D convolution has (1, 256, 40, 20)-kernel with no pooling. This corresponds to a frame length of 5 ms and a 2.5 ms shift. In the inference module, BLSTM layers have 3 layers with 256 nodes in each direction. Two LSTM-based decoders both run with 3 layers with 512 nodes. The dimension of the speaker and the direction embedding is 256. In the separation module, TCN runs with four convolution blocks and $R = 8$ in each block. Transposed convolution runs with (256, 1, 40, 20)-kernel. For loss, $\lambda = 5$. For SDNet, the input is raw-waveform, and it outputs raw-waveform.

3.2. Baselines

In our experiments, we build several baselines. SNet [13] acts as the baseline and is performed in the frequency domain. SNet-2ch represents the dual-channel version of SNet. We also build a dual-channel TasNet, named TasNet-2ch, whose channel differences are

Table 2: *The effect of different configurations on dual-channel datasets and comparisons on SISNR and SDR improvement (dB).*

System	Domain	Data Type	WSJ0-2mix		WSJ0-3mix		WSJ0-2&3mix	
			SISNRi	SDRi	SISNRi	SDRi	SISNRi	SDRi
SNet-2ch	Freq.	Anechoic	14.62	14.25	10.31	10.03	11.12	11.02
SNet-time-2ch	Time	Anechoic	20.88	20.61	14.32	14.11	17.43	17.31
SNet-time-2ch+IAC	Time	Anechoic	21.13	20.89	15.41	15.02	18.65	18.11
SDNet	Time	Anechoic	25.71	25.31	17.46	17.26	21.92	21.56
TasNet-2ch	Time	Anechoic	25.21	25.08	17.31	17.06	—	—
SNet-2ch	Freq.	Reverberant	7.32	7.28	5.53	5.15	6.53	6.33
SNet-time-2ch	Time	Reverberant	8.43	8.35	6.62	6.41	7.32	7.30
SNet-time-2ch+IAC	Time	Reverberant	8.76	8.59	6.93	6.86	7.88	7.64
SDNet	Time	Reverberant	10.57	10.64	8.49	8.55	9.91	9.08
TasNet-2ch	Time	Reverberant	10.78	10.83	9.08	9.32	—	—

learned in an end-to-end manner. These models are trained with the same datasets as our models.

3.3. Analysis of the proposed methods

Learned from the experimental results in Table 1 and Table 2, the proposed methods can effectively separate the mixed speech. In Table 1, SNet is first transferred into the time domain as SNet-time. Compared with SNet, SNet-time achieves performance improvement, which attributes to the time-domain-based end-to-end training. SNet-time-2ch means the dual-channel SNet-time. Compared with SNet-time, SNet-time-2ch achieves a significant performance improvement. It means that the spatial information can be utilized by our network to improve the performance.

IAC is used to extract the differences between channels. The extracted features are only used in the inference module. The time-domain-based dual-channel model with IAC is named as SNet-time-2ch+IAC. As shown in Table 2, the models with IAC have achieved performance improvement both on the anechoic and reverberant datasets.

When reverberation is added, performance is degraded. SDNet has achieved performance improvements both on anechoic and reverberant datasets. When separating the mixture, the speaker and direction can be inferred by SDNet. The inferred speaker and direction information is conducive to the selection of output. Our final model, SDNet, can achieve SDR improvements of 25.31 dB, 17.26 dB, and 21.56 dB on the anechoic WSJ0-2mix, WSJ0-3mix, and WSJ0-2&3mix datasets and 10.64 dB, 8.55 dB, and 9.08 dB on the reverberant WSJ0-2mix, WSJ0-3mix, and WSJ0-2&3mix datasets, respectively.

The performances of SNet-time and SNet-time-2ch are worse than the corresponding TasNets. This is due to the speaker mismatch between the training set and the test set, resulting in inaccurate speaker masks generated during the test. The direction inference mechanism in SDNet can effectively alleviate this problem. Reverberation has a negative impact on direction inference. SDNet performs similar to TasNet-2ch, but it does not need prior knowledge of the number of outputs.

3.4. Inference accuracy of sound-source number

The advantage of our proposed model is that it can dynamically estimate the number of sound sources. In the WSJ0-2mix and WSJ0-3mix, the number of speakers mixed in speech is fixed. We find that the models learn this pattern. In these experiments, the inference

Table 3: *Inferring accuracy of source numbers on reverberant WSJ0-2&3mix dataset.*

Model	Accuracy (%)
SNet-time	81.75
SNet-time-2ch	85.11
SNet-time-2ch+IAC	86.67
SDNet	89.73

accuracies are close to 100%. Therefore, we construct the WSJ0-2&3mix dataset and perform experiments on this dataset. The experimental results are shown in Table 3.

For comparison, SNet-time in Table 3 is performed on the same reverberant dataset but only on the reference channel. In Table 3, compared with SNet-time, the inference accuracy of SNet-time-2ch has been greatly improved, which indicates that the spatial information is learned by the model and used to increases the discrimination of the sound sources. Compared with SNet-time-2ch, SNet-time-2ch+IAC can infer the number of sound sources more accurately. This shows that the extracted channel differences are beneficial to our system. SDNet achieves 89.73%, which indicates the proposed method can make better use of the spatial information.

4. CONCLUSIONS

We propose a time-domain-based speaker and direction-inferred dual-channel speech separation network, which first infers the speaker with direction and then integrates them as a source mask to separate the mixed speech. Experimental results show that SDNet effectively separates mixture under anechoic and reverberant conditions and deals with the problem of an unknown number of sources in the mixture and selection of outputs.

5. ACKNOWLEDGMENTS

6. REFERENCES

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

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

[3] 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.

[4] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *IEEE ICASSP*, 2018, pp. 686–690.

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

[6] Z.-Q. Wang, K. Tan, and D. Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," in *IEEE ICASSP*, 2019, pp. 71–75.

[7] Y. Liu and D. Wang, "Divide and conquer: A deep casa approach to talker-independent monaural speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2092–2102, 2019.

[8] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[9] Z. Shi, H. Lin, L. Liu, R. Liu, J. Han, and A. Shi, "Deep attention gated dilated temporal convolutional networks with intra-parallel convolutional modules for end-to-end monaural speech separation," in *Interspeech*, 2019, pp. 3183–3187.

[10] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *IEEE ICASSP*, 2020, pp. 46–50.

[11] T. Higuchi, K. Kinoshita, M. Delcroix, K. Zmolíková, and T. Nakatani, "Deep clustering-based beamforming for separation with unknown number of sources." in *Interspeech*, 2017, pp. 1183–1187.

[12] N. Takahashi, S. Parthasaarathy, N. Goswami, and Y. Mitsufuji, "Recursive speech separation for unknown number of speakers," in *Interspeech*, 2019, pp. 1348–1352.

[13] J. Shi, J. Xu, and B. Xu, "Which ones are speaking? speaker-inferred model for multi-talker speech separation," in *Interspeech*, 2019, pp. 4609–4613.

[14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations*, 2015.

[15] P. Yang, X. Sun, W. Li, S. Ma, W. Wu, and H. Wang, "Sgm: Sequence generation model for multi-label classification," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 3915–3926.

[16] Y. Luo and N. Mesgarani, "Separating varying numbers of sources with auxiliary autoencoding loss," in *Interspeech*, 2020.

[17] 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," in *Interspeech*, 2018, pp. 307–311.

[18] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *IEEE ICASSP*, 2018, pp. 5554–5558.

[19] 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 *Interspeech*, 2019, pp. 2728–2732.

[20] C. Xu, W. Rao, E. S. Chng, and H. Li, "Spex: Multi-scale time domain speaker extraction network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1370–1384, 2020.

[21] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network," in *Interspeech*, 2020.

[22] 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 *Interspeech*, 2019, pp. 2713–2717.

[23] R. Gu and Y. Zou, "Temporal-spatial neural filter: Direction informed end-to-end multi-channel target speech separation," *arXiv preprint arXiv:2001.00391*, 2020.

[24] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *IEEE Conference on Computer Vision and Pattern ecognition*, 2017, pp. 1251–1258.

[25] Z.-Q. Wang and D. Wang, "Integrating spectral and spatial features for multi-channel speaker separation." in *Interspeech*, 2018, pp. 2718–2722.

[26] R. Gu, J. Wu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "End-to-end multi-channel speech separation," *arXiv preprint arXiv:1905.06286*, 2019.

[27] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[29] S. Wiseman and A. M. Rush, "Sequence-to-sequence learning as beam-search optimization," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1296–1306.

[30] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[31] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Interspeech*, 2016, pp. 545–549.

[32] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[33] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.