SAGRNN: Self-Attentive Gated RNN For Binaural Speaker Separation With Interaural Cue Preservation

Ke Tan [ID], Buye Xu [ID], Anurag Kumar, Eliya Nachmani, and Yossi Adi

Abstract—**Most existing deep learning based binaural speaker separation systems focus on producing a monaural estimate for each of the target speakers, and thus do not preserve the interaural cues, which are crucial for human listeners to perform sound localization and lateralization. In this study, we address talker-independent binaural speaker separation with interaural cues preserved in the estimated binaural signals. Specifically, we extend a newly-developed gated recurrent neural network for monaural separation by additionally incorporating self-attention mechanisms and dense connectivity. We develop an end-to-end multiple-input multiple-output system, which directly maps from the binaural waveform of the mixture to those of the speech signals. The experimental results show that our proposed approach achieves significantly better separation performance than a recent binaural separation approach. In addition, our approach effectively preserves the interaural cues, which improves the accuracy of sound localization.**

Index Terms—**Binaural speaker separation, self-attention, interaural cue preservation, time-domain.**

I. INTRODUCTION

IN REAL acoustic environments, a speech source of interest is frequently corrupted by interfering sounds. Human auditory system excels at attending to a target speech source, and the cocktail party problem [5] aims to develop such capabilities in man-made devices and systems. A critical aspect of the cocktail party problem is speaker separation which aims to separate multiple concurrent speech signals of interest from a sound mixture.

Conventionally, most of the speaker separation methods work in time-frequency (T-F) domain where T-F representations are typically computed using short-time Fourier transform (STFT). In recent years, the performance of T-F domain speaker separation has substantially improved due to the use of deep learning [4], [13], [15], [18], [20], [34]–[36], [38]. Moreover, the advent of deep learning based speech separation has also ignited interest in time-domain approaches, which directly estimate the waveform of clean speech from that of the mixture without

resorting to a T-F representation. A notable time-domain speaker separation approach is TasNet [22], which yields comparable scale-invariant signal-to-noise ratios (SI-SNRs) and signal-to-distortion ratios (SDRs) to the ideal ratio mask (IRM). Other related studies include [21], [26], [30], [33], [40] and [39].

While several time-domain monaural speaker separation methods have been developed, very few works have focused on binaural separation. Moreover, most existing binaural separation systems have a multiple-input single-output (MISO) layout, which produce a mono estimate for each of the target speakers from a binaural mixture [6], [19], [42]. Hence these systems do not preserve interaural cues such as interaural time differences (ITDs) and interaural level differences (ILDs), which are crucial for human listeners to perform sound localization and lateralization [7], [14].

On the T-F domain front, various techniques have been developed to preserve binaural cues in the estimated signals. One can apply a common real-valued T-F mask or spectral gain to both left and right channels [23], [43]. Alternatively, binaural cues can be preserved by applying adaptive beamformers with additional constraints that encourage interaural cue preservation [10], [24], [31]. However, these techniques sacrifice the separation performance. More recently, a multiple-input multiple-output (MIMO) TasNet [11] was designed, which produces a binaural estimate for each speaker. MIMO TasNet yields significantly better speech quality than the single-channel TasNet while preserving both ITDs and ILDs.

In this letter, we propose a novel framework called *multiple-input multiple-output self-attentive gated recurrent neural network* (MIMO SAGRNN) for binaural speaker separation. The proposed SAGRNN network architecture extends the gated RNN in [26] by additionally incorporating self-attention mechanisms and dense connectivity (DC). We then derive MIMO SAGRNN from a single-input single-output (SISO) SAGRNN by first extending the SISO SAGRNN into a multiple-input single-output (MISO) layout by creating two encoders, one for the *reference ear* input and the other for the *non-reference ear* input. This MISO SAGRNN estimates the separated signals in the reference ear. The MIMO system is formulated by alternately treating each ear as the reference ear, yielding estimates for both ears in a symmetric manner.

The rest of this letter is organized as follows. Section II describes our proposed approach. The experimental results are presented in Section III, and Section IV concludes this letter.

II. ALGORITHM DESCRIPTION

We progressively develop a MIMO system for binaural speaker separation. Specifically, we start with a SISO SAGRNN architecture, and then present the MIMO setup.

Fig. 1. (Color Online). Diagram of SISO SAGRNN. The dotted lines represent the procedures that exist only during training.
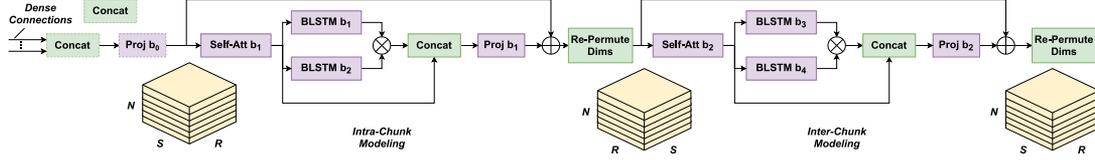
Fig. 2. (Color Online). Diagram of the SA-MULCAT block. The dotted lines indicate that the procedures exist in all SA-MULCAT blocks except the first one. The symbol ⊗ represents the element-wise multiplication, and ⊕ the element-wise addition.

A. SISO SAGRNN

As in [26], the separation framework of a SISO SAGRNN comprises three stages: *encoding and chunking*, *block processing*, and *decoding and overlap-add*. A time-domain input mixture is transformed into a set of overlapped chunks via encoding and chunking, which leads to a 3-D embedding. Subsequently, the 3-D embedding is passed into stacked RNN blocks to perform intra-chunk (local) and inter-chunk (global) modeling alternately and iteratively. The 3-D representation learned by the last RNN block is decoded and then transformed back to the time domain by an overlap-add operator.

Given a T-sample input waveform $y \in \mathbb{R}^T$, an encoder is used to segment and encode y into L overlapped time frames with a frame size of P and a hop size of $P/2$, yielding a 2-D embedding $\mathbf{U} \in \mathbb{R}^{N \times L}$. Specifically, the encoder consists of a 1-D strided convolutional layer with N output channels, followed by a rectified linear activation function. We divide the time frames into S overlapped chunks with a chunk size of R and a hop size of $R/2$. These chunks are then concatenated into a 3-D embedding $\widetilde{\mathbf{W}} = [\mathbf{W}_1, \ldots, \mathbf{W}_S] \in \mathbb{R}^{N \times S \times R}$, where $\mathbf{W}_1, \ldots, \mathbf{W}_S \in \mathbb{R}^{N \times R}$ are the 2-D chunks.

Subsequently, the 3-D embedding $\widetilde{\mathbf{W}}$ is fed into a series of B RNN blocks for processing. To improve the information and gradient flow between blocks, we propose a dense connectivity pattern: each block receives the outputs of all preceding blocks, i.e. $\widetilde{\mathbf{W}}_b = \mathcal{H}_b([\widetilde{\mathbf{W}}_0, \ldots, \widetilde{\mathbf{W}}_{b-1}])$ for $b = 1, \ldots, B$,

where $\mathcal{H}_b$ denotes the mapping function defined by the b-th block, and $[\cdot, \ldots, \cdot]$ the concatenation operation. The output embedding of the b-th block is represented by $\widetilde{\mathbf{W}}_b$, where $\widetilde{\mathbf{W}}_0 = \widetilde{\mathbf{W}}$ and $\widetilde{\mathbf{W}}_b \in \mathbb{R}^{N \times S \times R}, \forall b$. The dense connections encourage feature reuse among blocks, which explicitly leverage different information learned by different blocks.

Similar to [26], we use a multi-scale loss for training, which necessitates producing a waveform estimate for each speaker after each block. We decode the output embedding of each block with a decoder, which comprises a parametric rectified linear function [12] followed by a 2-D 1×1 convolutional layer with $C \cdot N$ output channels. The decoded feature of size $CN \times S \times R$ is divided into C 3-D representations of size $N \times S \times R$, corresponding to the C speech sources. These
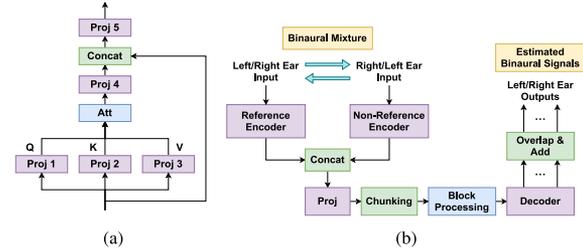
Fig. 3. (Color Online). (a) Diagram of the self-attention block. (b) Overview of the MIMO separation system with an underlying MISO system.

3-D representations are transformed back to waveforms by two successive overlap-add operations at the chunk level and the frame level, respectively. Note that the same decoder is applied to the output of each block. Fig. 1 depicts the SISO SAGRNN.

A series of multiply-and-concatenate (MULCAT) blocks are employed to model the intra-chunk and inter-chunk dependencies. In this study, we extend the MULCAT block by introducing self-attention [32], which amounts to a self-attention based MULCAT (SA-MULCAT) block illustrated in Fig. 2. The concatenation of paths from the dense connections is fed into a linear projection layer for dimension reduction, yielding an embedding of size $N \times S \times R$. The resulting embedding is successively passed through two subblocks, one for intra-chunk modeling and the other for inter-chunk modeling. In each subblock, we employ a self-attention block followed by a gated RNN module, which consists of two bidirectional long short-term memory (BLSTM) layers coupled with each other. Each BLSTM contains H units in each direction. The Hadamard product of their outputs is concatenated with the input to the gated RNN module, and then passed into a linear projection layer for dimension reduction. In addition, a skip connection is used to bypass the subblock. After the first subblock, the dimensions of the 3-D representation are re-permuted, so that sequential modeling can be performed across chunks in the second subblock. After the second subblock, the dimensions are re-permuted back.

The self-attention block is illustrated in Fig. 3(a). We first divide a 3-D representation into a set of 2-D slices $\mathbf{Z} \in \mathbb{R}^{M \times N}$, where $M = R$ for intra-chunk modeling and $M = S$ for inter-chunk modeling. Each slice is linearly projected to a query

matrix $\mathbf{Q}$, a key matrix $\mathbf{K}$ and a value matrix $\mathbf{V}$ by three different projection layers, where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{M \times D}$ and D is set to 64. We apply a scaled dot-product attention function:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = SoftMax\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{D}}\right)\mathbf{V}, \quad (1)$$

where $SoftMax(\cdot)$ denotes the softmax function across columns. The output of the attention function is computed as a weighted sum of the values, where the weight assigned to each value is derived by measuring the similarities between the queries and the keys. Subsequently, all the attention output slices are merged and then linearly projected back to the size of the input 3-D representation. With a skip connection, this representation is concatenated with the input to the self-attention block, and then projected back to the original size. The use of self-attention is motivated by its recent success on monaural speech enhancement and dereverberation [17], [41], which has demonstrated its capability of capturing long-term dependencies in target speech and interference. By leveraging the relevance among features at different time steps, self-attention produces a dynamic representation in adapting to different acoustic conditions.

B. MIMO SAGRNN

As shown in Fig. 3(b), a *reference encoder* and a *non-reference encoder* are employed to process the binaural mixture waveforms. The resulting 2-D embeddings are concatenated and then linearly projected to the size of $N \times L$. Subsequently, we successively perform block processing, decoding and overlap-add, akin to the SISO system. In this MISO system, the separation outputs always correspond to the reference ear. We formulate the MIMO system by alternately treating each ear as the reference. Specifically, the separation outputs for the left ear are obtained by treating the left ear as the reference ear and the right ear as the non-reference. The separation outputs for the right ear are obtained by swapping the inputs of the two ears. Note that the same MISO system is used for separation in both channels. Such a cross-ear referencing strategy selects the target channel by exploiting discriminative information within the ordered pair of channels.

C. Training Objective

We use the plain SNR rather than the widely-used SI-SNR [22] as the training objective. The rationale is that SI-SNR training cannot preserve the ILD in the binaural estimates, as the power scale of the estimated signals is insusceptible to training due to the scale invariance. The mean of the SNR losses from all SA-MULCAT blocks is used for training. The waveforms of the clean speech signals are used as the training target for calculating the losses from all blocks. In addition, we apply the permutation invariant training [18] criterion to the loss from each block individually, which allows the label permutation to change from one block to another.

III. EXPERIMENTS

A. Experimental Setup

We simulate a noise-free dataset and a noisy dataset from the WSJ0-2mix dataset [9], [13], which contains 20,000, 5,000 and 3,000 mixtures in the training, validation and testing sets,

TABLE I
COMPARISON OF DIFFERENT SYSTEMS IN THE NOISE-FREE CONDITION

Metrics	# param.	ΔSDR (dB)	ΔSNR (dB)	ESTOI (%)	PESQ
Mixture	-	0.00	0.00	56.10	1.99
TasNet	6.66 M	15.78	15.80	90.25	3.30
DPRNN	6.99 M	18.25	18.28	92.64	3.49
Gated RNN	7.56 M	18.37	18.41	92.52	3.52
Oracle IBM	-	13.69	13.61	88.49	3.36
Oracle IRM	-	13.05	13.07	93.33	3.73
Oracle PSM	-	16.77	16.58	95.55	3.91
MIMO TasNet	7.32 M	21.14	20.69	95.53	3.73
Oracle MB-MVDR	-	17.13	10.44	95.77	3.66
MIMO SAGRNN	8.71 M	**27.19**	**26.88**	**98.08**	**4.06**
− half multi-scale loss (i)	8.71 M	24.35	23.99	96.70	3.90
− multi-scale loss (ii)	8.71 M	22.93	22.45	95.69	3.79
− DC (iii)	8.38 M	24.02	23.64	96.54	3.87
− SA (iv)	7.92 M	23.13	22.75	96.12	3.82
− DC − SA (v)	7.59 M	21.97	21.64	95.66	3.73

respectively. For both datasets, we convolve each pair of utterances in WSJ0-2mix with two randomly sampled head-related impulse responses (HRIRs) from the CIPIC HRTF Database [3] respectively, which contains 45 subjects with 25 (azimuths) × 50 (elevations) directions for each subject. Specifically, we choose 35 subjects for training and cross validation, and use the 10 remaining subjects for testing. For the noisy dataset, we additionally simulate uncorrelated noise sources by randomly selecting HRIRs for them, where the number of noise sources is randomly sampled between 1 and 10. Note that all sound sources are placed in different directions. We use a set of roughly 65,000 noises from the DNS Challenge [28] for training and cross validation, and a different set [2] of roughly 1,300 noises for testing. The SNR (w.r.t. the speech mixture in the left ear) is randomly chosen between -10 dB and 10 dB. All signals are sampled at 8 kHz.

We train the models on 4-second segments with the AMSGrad optimizer [27] with a minibatch size of 4. The learning rate is initialized to 0.0002, which decays by 0.98 every 2 epochs. Gradient clipping with a maximum ℓ^2 norm of 3 is applied during training. The network hyperparameters for MIMO SAGRNN are as follow: $P = 8$, $N = 128$, $R = 126$, $H = 128$, $D = 64$ and $B = 6$. Note that the value of R is selected such that $R \approx S = 128$ for the training segments.

We use several monaural and binaural separation models for comparison; the monaural models are trained and evaluated on each ear individually. Specifically, we use TasNet [22], dual-path RNN (DPRNN) [21] and the gated RNN in [26] as monaural baselines. We use MIMO TasNet [11] as a binaural baseline. We slightly adjust the hyperparameter configurations of all baselines, so that they have comparable model sizes to our MIMO SAGRNN. For the noncausal temporal convolutional network (TCN) in TasNet [22], the number of repeated stacks is set to 4. For DPRNN, the number of output channels in the encoder and the decoder is set to 128, and the number of units in each direction for each BLSTM to 200. For MIMO TasNet, we replace the causal TCN by a noncausal TCN, with bottleneck size of 128. In addition, the number of output channels in the encoder and the decoder of MIMO TasNet is set to 512.

B. Experimental Results

1) Separation Results and Analysis: Tables I and II show comparison among different approaches for the noise-free and noisy conditions respectively. The separation results are reported in terms of SDR improvement (ΔSDR), SNR improvement (ΔSNR), extended short-time objective intelligibility (ESTOI) [16], and perceptual evaluation of speech quality (PESQ) [29]. We can observe that MIMO TasNet produces consistently better results than the monaural baselines. Moreover, our proposed MIMO SAGRNN substantially outperforms

TABLE II
COMPARISON OF DIFFERENT SYSTEMS IN THE NOISY CONDITION

Metrics	ΔSDR (dB)	ΔSNR (dB)	ESTOI (%)	PESQ
Mixture	0.00	0.00	27.80	1.49
TasNet	11.71	13.33	54.65	2.12
DPRNN	11.91	13.55	53.72	2.17
Gated RNN	12.89	14.27	59.09	2.30
Oracle IBM	14.11	14.64	71.19	2.63
Oracle IRM	13.19	13.97	84.22	3.33
Oracle PSM	17.41	17.62	89.40	3.55
MIMO TasNet	14.40	15.23	63.79	2.41
Oracle MB-MVDR	4.98	4.90	42.71	1.79
MIMO SAGRNN	**17.53**	**17.95**	**75.14**	**2.78**

TABLE III
COMPARISON BETWEEN SISO AND MIMO SAGRNNS ON THE BETTER-EAR
CHANNEL FOR THE NOISY CONDITION

Metrics	ΔSDR (dB)	ΔSNR (dB)	ESTOI (%)	PESQ
Mixture	0.00	0.00	31.84	1.57
Gated RNN	12.59	13.50	64.64	2.42
SISO SAGRNN	13.19	14.06	66.91	2.49
MIMO SAGRNN	15.94	16.32	75.66	2.79

MIMO TasNet in all the four metrics. For the noise-free condition, MIMO SAGRNN improves SDR by 6.05 dB and SNR by 6.19 dB over MIMO TasNet. Some demos can be found at https://jupiterethan.github.io/sagrnn.github.io/.

In addition, we compare MIMO SAGRNN with several oracle approaches, including ideal binary mask (IBM), ideal ratio mask (IRM), phase-sensitive mask (PSM) [8] and an oracle masking-based minimum variance distortionless response (MB-MVDR) beamformer. We use an open-source implementation [1] of the oracle MB-MVDR beamformer, with a frame length of 64 ms and a frame shift of 32 ms. The IRM is used to calculate the spatial covariance matrices. We alternately treat each channel as the reference channel to produce the binaural estimate. As shown in Table I, our approach consistently outperforms the ideal masks and the oracle beamformer in the noise-free condition. In the noisy condition (Table II), our approach produces slightly higher SDR and SNR but lower ESTOI and PESQ than the PSM. Note that, ESTOI and PESQ improvements over the mixtures using the oracle MB-MVDR beamformer dramatically decrease in the noisy condition compared with the noise-free condition. This is likely because the directionality of the sound sources is smeared due to the presence of multiple noise sources. In contrast, our approach is more robust against the noise field.

2) Ablation Study: We conduct an ablation study to understand the contribution of each component in our approach. Several variants of MIMO SAGRNN are compared in Table I: (i) using multi-scale loss computed from only the last three RNN blocks; (ii) using loss computed from only the last RNN block; (iii) without dense connections; (iv) without self-attention blocks; (v) without dense connections and self-attention blocks. It is shown that self-attention and dense connectivity are crucial for MIMO SAGRNN. Without self-attention, for example, SDR decreases by 4.06 dB and SNR by 4.13 dB. We also compare the SISO and MIMO SAGRNNs on the better-ear channel; *better ear* is defined as the ear that is closer to the target speech source. The azimuth position of the speech source is used to determine the better ear. Table III shows that MIMO SAGRNN yields significantly better results than SISO SAGRNN and the gated RNN in [26] on all metrics, evidencing that the binaural inputs are effectively leveraged by our MIMO system.

3) Evaluation of Interaural Cue Preservation: Lastly, we evaluate the preservation of interaural cues in the estimated binaural signals under noise-free condition. We apply a binaural sound localization algorithm [25] to the binaural estimates, of which an open-source implementation is available. This implementation estimates the azimuth position of the sound source

TABLE IV
EVALUATION OF INTERAURAL CUE PRESERVATION WITH BINAURAL SOUND
LOCALIZATION FOR THE NOISE-FREE CONDITION

Metrics	ΔAzimuth (°)	ΔITD (μs)	ΔILD (dB)		
Frequency Channels	-	-	2.07 kHz	3.08 kHz	3.75 kHz
Mixture	26.03	255.85	4.31	4.85	5.12
TasNet	14.40	25.71	0.62	0.68	0.95
DPRNN	13.64	25.71	0.62	0.68	0.93
Gated RNN	13.82	23.42	0.91	0.73	1.22
Oracle IBM	12.25	15.75	0.90	0.98	1.16
Oracle IRM	9.43	55.00	0.29	0.28	0.39
Oracle PSM	4.95	16.20	0.37	0.40	0.55
MIMO TasNet	6.45	20.10	0.88	0.77	1.13
Oracle MB-MVDR	34.05	371.35	6.31	7.06	7.39
MIMO SAGRNN	**5.88**	**14.95**	**0.53**	**0.45**	**0.70**

(a) ITD Dist. (Clean)   (b) ITD Dist. (Estimated)

(c) ILD Dist. at 3.75 kHz (Clean)   (d) ILD Dist. at 3.75 kHz (Estimated)
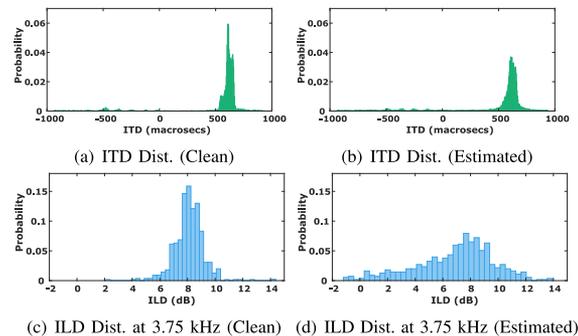
Fig. 4.   (Color Online). An example of ITD and ILD distributions.

at the frame level, as well as the ITD and the ILD for each T-F unit of a 32-channel cochleagram based on a gammatone filter-bank. The average frame-level azimuth errors are presented in Table IV. Given the dominance of the ITD cue at low frequencies (below 1.5 kHz) in sound localization [37], we only take into account the frequency bands corresponding to gammatone filters with a maximum center frequency of approximately 1.5 kHz. Since ILD is highly frequency-dependent due to diffraction and attenuation of the sounds, we calculate the average ILD errors individually for three empirically selected frequency channels, corresponding to the gammatone filters with the center frequencies of roughly 2.07, 3.08 and 3.75 kHz. Given the fact that all sound sources are stationary in this study, we summarize only one ITD/ILD from an entire utterance in the following way. We plot a histogram of the T-F unit level ITDs/ILDs, and then estimate the ITD/ILD based on the center value of the highest bin. The number of bins is empirically set to 500 for ITD and 40 for ILD. An example of ITD and ILD histograms are presented in Fig. 4. As shown in Table IV, our approach reduces the azimuth error by 20.15° compared to the mixtures. Moreover, our approach yields consistently smaller azimuth, ITD and ILD errors than MIMO TasNet and the monaural baselines, showing that our approach preserves the interaural cues more effectively.

IV. CONCLUSION

We have proposed an end-to-end MIMO system for binaural speaker separation with interaural cue preservation. We developed a novel framework which relies on self-attention and dense connectivity for improved speaker separation. Our experimental results show that the proposed approach significantly outperforms a binaural separation approach (i.e. MIMO TasNet) in terms of ΔSDR, ΔSNR, ESTOI and PESQ. Moreover, our approach effectively preserves the auditory spatial cues of talkers. For future work, we would devote more efforts to the design of MIMO systems for real-time processing, as well as exploring binaural speaker separation in more realistic acoustic conditions (e.g. with reverberation and diffuse noise).

REFERENCES

[1] Beamformers. 2020. [Online]. Available: https://pypi.org/project/beamformers

[2] Robinhood76 sounds. 2008. [Online]. Available: https://freesound.org/people/Robinhood76/

[3] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2001, pp. 99–102.

[4] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 246–250.

[5] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975–979, 1953.

[6] P. Dadvar and M. Geravanchizadeh, "Robust binaural speech separation in adverse conditions based on deep neural network with modified spatial features and training target," *Speech Commun.*, vol. 108, pp. 41–52, Apr. 2019.

[7] R. Domnitz and H. Colburn, "Lateral position and interaural discrimination," *J. Acoust. Soc. Amer.*, vol. 61, no. 6, pp. 1586–1598, 1977.

[8] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, " Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 708–712.

[9] J. Garofolo, D. Graff, D. Paul, and D. Pallett, CSR-I (WSJ0) complete LDC93S6A. Web Download. Philadelphia, PA, USA: Linguistic Data Consortium, 83, 1993.

[10] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2449–2464, Dec. 2015.

[11] C. Han, Y. Luo, and N. Mesgarani, "Real-time binaural speech separation with preserved spatial cues," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6404–6408.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[13] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 31–35.

[14] R. Hershkowitz and N. Durlach, "Interaural time and amplitude jnds for a 500-Hz tone," *J. Acoust. Soc. Amer.*, vol. 46, no. 6B, pp. 1464–1467, 1969.

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

[16] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[17] Y. Koizumi, K. Yaiabe, M. Delcroix, Y. Maxuxama, and D. Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 181–185.

[18] 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. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[19] Q. Liu, Y. Xu, P. J. Jackson, W. Wang, and P. Coleman, "Iterative deep neural networks for speaker-independent binaural blind speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 541–545.

[20] Y. Liu and D. L. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 27, no. 12, pp. 2092–2102, Dec. 2019.

[21] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 46–50.

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

[23] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[24] D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, "Theoretical analysis of linearly constrained multi-channel Wiener filtering algorithms for combined noise reduction and binaural cue preservation in binaural hearing aids," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 23, no. 12, pp. 2384–2397, Dec. 2015.

[25] T. May, S. Van De Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 1–13, Jan. 2011. [Online]. Available: http://amtoolbox.sourceforge.net/amt-0.10.0/doc/models/may2011.php

[26] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *Proc. Int. Conf. Mach. Learn.*, 2020, vol. 119, pp. 7164–7175.

[27] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of Adam and beyond," in *Proc. Int. Conf. Learn. Representations*, 2018, *arXiv:1904.09237*.

[28] C. K. Reddy *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," 2020, *arXiv:2005.13981*.

[29] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2001, vol. 2, pp. 749–752.

[30] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *Proc. 19th Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 334–340.

[31] T. Van den Bogaert, J. Wouters, S. Doclo, and M. Moonen, "Binaural cue preservation for hearing aids using an interaural transfer function multichannel Wiener filter," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2007, vol. 4, pp. 565–568.

[32] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[33] S. Venkataramani, J. Casebeer, and P. Smaragdis, "End-to-end source separation with adaptive front-ends," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, 2018, pp. 684–688.

[34] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 686–690.

[35] Z.-Q. Wang, J. Le Roux, D. L. Wang, and J. R. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," in *Proc. Interspeech*, 2018, pp. 2708–2711.

[36] Z.-Q. Wang, K. Tan, and D. L. Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 71–75.

[37] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1648–1661, Mar. 1992.

[38] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 241–245.

[39] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," 2020, *arXiv:2002.08933*.

[40] L. Zhang, Z. Shi, J. Han, A. Shi, and D. Ma, "FurcaNeXt: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks," in *Proc. Int. Conf. Multimedia Model.*, 2020, pp. 653–665.

[41] Y. Zhao, D. L. Wang, B. Xu, and T. Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1598–1607, 2020.

[42] M. Zohourian and R. Martin, "Binaural speaker localization and separation based on a joint ITD/ILD model and head movement tracking," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 430–434.

[43] M. Zohourian and R. Martin, "GSC-based binaural speaker separation preserving spatial cues," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 516–520.