VARARRAY: ARRAY-GEOMETRY-AGNOSTIC CONTINUOUS SPEECH SEPARATION

Takuya Yoshioka, Xiaofei Wang, Dongmei Wang, Min Tang, Zirun Zhu, Zhuo Chen, Naoyuki Kanda

Microsoft, One Microsoft Way, Redmond, WA, USA

ABSTRACT

Continuous speech separation using a microphone array was shown to be promising in dealing with the speech overlap problem in natural conversation transcription. This paper proposes VarArray, an array-geometry-agnostic speech separation neural network model. The proposed model is applicable to any number of microphones without retraining while leveraging the nonlinear correlation between the input channels. The proposed method adapts different elements that were proposed before separately, including transform-average-concatenate, conformer speech separation, and inter-channel phase differences, and combines them in an efficient and cohesive way. Large-scale evaluation was performed with two real meeting transcription tasks by using a fully developed transcription system requiring no prior knowledge such as reference segmentations, which allowed us to measure the impact that the continuous speech separation system could have in realistic settings. The proposed model outperformed a previous approach to array-geometry-agnostic modeling for all of the geometry configurations considered, achieving asclite-based speaker-agnostic word error rates of 17.5% and 20.4% for the AMI development and evaluation sets, respectively, in the end-to-end setting using no ground-truth segmentations.

Index Terms— Speech separation, microphone arrays, array-geometry-agnostic modeling, meeting transcription

1. INTRODUCTION

The last decade has witnessed transformational progress in automatic speech recognition (ASR) technology driven by deep learning approaches. Modern ASR systems can accurately transcribe pre-segmented utterances recorded in acoustically moderate environments, as demonstrated in many ASR benchmarks such as LibriSpeech [1] and Switchboard [2], and are widely used in our daily lives. Yet, there are many challenges that must be overcome for the ASR technology to become usable more broadly [3, 4]. One of the unsolved challenges is transcribing unsegmented natural conversations, where multiple people talk over each other in a spontaneous and thus unpredictable way.

Continuous speech separation (CSS) was proposed to handle the speech overlaps in natural human-to-human conversation transcription [5–7]. With this approach, continuously captured long-form conversational signals are split into multiple overlap-free signals. Since each of the output signal does not contain the speech overlaps, they can be processed with conventional ASR systems. A microphone array-based approach has been particularly successful and was applied to real meeting recordings [6] while most previous studies used artificially created meeting-like data.

This paper sheds light on the generalizability of the multi-channel speech separation models. Two major approaches exist for the multi-channel neural network modeling for speech separation. One approach uses multi-channel features confined to a specific microphone array geometry [8, 9]. While it can potentially utilize the full spatial information, the trained model can be used only for the microphone array that it is trained for. Another approach is to apply the same speech separation model to each input channel and merge the outputs from the individual channels by calculating their median or average [10, 11]. While being applicable to any microphone array devices, this approach does not take advantage of the nonlinear correlation between the channels. Also, the computational cost becomes too large to deploy in practice when the number of channels is large.

This problem is addressed by VarArray, our proposed array geometry-agnostic speech separation model. The input is a set of magnitude and inter-channel phase difference (IPD) features, organized in a way that is invariant to the microphone permutation. The model consists of conformer blocks [12, 13] interleaved with transform-average-concatenate (TAC) layers [14] to exploit the spatio-temporal patterns exhibited in the input. TAC is a cross-channel layer model that can cope with any number of channels in a permutation-invariant fashion. To reduce the growth rate of the model evaluation cost with respect to the number of microphones, the multiple feature streams are merged early in the network.

Large-scale evaluation was carried out by using real meeting recordings. Two meeting corpora were used to evaluate the performance sensitivity to the array geometry: AMI [15] and Microsoft internal meetings [6]. Our meeting transcription system achieved state-of-the-art results in AMI under the condition of no ground-truth segmentations being used. Results of fine-tuning the speech separation model to AMI with an ASR-based loss function are also presented, demonstrating its impact on the full-blown system. Note that, while the array-geometry-agnostic modeling was studied before in related areas, such as dereverberation [16], noise reduction [17], and ad hoc microphone arrays [18], this paper is the first to combine different pieces including the IPD features, conformer speech separation, and TAC in a cohesive and efficient way and presents thorough evaluation results for the real meeting tasks.

2. CONTINUOUS SPEECH SEPARATION

2.1. General architecture

CSS is a front-end-based approach to the speech overlap problem in natural human-to-human conversation transcription. Given long-form audio signals captured by a microphone array, the goal of CSS is to generate K signals that have the same length as the input in such a way that each of the K signals does not contain overlapped utterances inside while the sum of these K signals retains the spoken content present in the input signals in its entirety. It is tempting to try to estimate the clean speech signal of each speaker, which means K is equal to the total number of speakers. However, this approach is extremely difficult for streaming processing because it requires the front-end to perform speaker diarization in addition to enhancing the speech. A more practical approach, which is increasingly being adopted, is to address only the speech overlap problem [7, 9, 19, 20]. In meetings, the maximum number of simultaneously talking speakers is very limited. In fact, we can assume only two or fewer speakers to be active for the majority of the meeting time [21]. This means that only two output signals are sufficient in practice (i.e., $K = 2$). If the audio segment being processed has two overlapping utterances, we separate each utterance and send the obtained signals to different output channels. When the current segment consists of only one speaker, we just have to enhance the speech quality and route the processed signal to one of the output channels. The unused output channel can be filled by zero to make the two output signals synchronous. As the output signals are overlap-free, they can be directly fed to a conventional ASR system.

A sliding window-based approach is often adopted to enable CSS. At each window position, a speech separation model trained with utterance-level permutation invariant training (uPIT) [22] is applied to the windowed input signal to generate K separated signals. The order of the output signals is determined to maintain the consistency with the separated signals obtained at the previous window position. Specifically, we choose the output permutation that provides the smallest mean squared error between the pairs of the separated signals obtained at the current and previous window positions. The errors are calculated for the overlapping frames of the two windows. This “stitching processing” can be done by using the overlapped segment between the two adjacent windows. In the following, we focus our attention on the local speech separation and uPIT as it constitutes the central piece of the sliding window-based approach.

2.2. Local speech separation

The local speech separation process can be described as follows. Let X_{mft} denote the short time Fourier transform (STFT) coefficient of the audio signal observed by the mth microphone, where f and t represent the frequency and time indices, respectively. Feature vector $\boldsymbol{y}_t$ is computed for each time frame, and the sequence of the feature vectors is fed to a speech separation neural network model. The model produces time-frequency (TF) masks for predicting the two clean speech signals (recall that we assume $K = 2$) and the background noise. The masks are used to perform minimum variance distortionless response (MVDR) beamforming to estimate the clean signals. It was empirically found that sparcifying the TF masks before the MVDR computation improved the ASR accuracy. In our experiments, this was performed by retaining only the most dominant sound source for each TF bin. Gain adjustment of [8] was also applied after beamforming to counteract MVDR’s inability of generating complete silence for non-overlapped signals.

During training, reference signals are available for the clean speech and the noise. For training samples with one active speaker, one of the reference signals is a zero sequence. The uPIT loss [22] is employed to cope with the arbitrariness of the order of the two speech signals.

The feature vector, $\boldsymbol{y}_t$, can be the concatenation of the STFT coefficients of all the microphones [9] or a stack of IPDs [8], just to mention a few examples. Various speech separation models can be applied to these concatenated features, including conformer networks [13] and dual-path networks [23]. However, using the concatenated features make the system dependent on a particular microphone array and usable only with the audio device which it is trained for. Our goal is to create a speech separation model that can work with any microphone arrays which performs as well as or better than the array geometry-dependent ones.

3. VARARRAY

VarArray, our proposed speech separation model, can efficiently utilize the spatial and temporal information from any number of microphone inputs for accurate TF mask estimation. It also yields the same output regardless of the microphone permutation. There are two key ingredients: one concerns the feature design; the other relates to the model structure. They are described in Sections 3.1 and 3.2, respectively. In the following, $\mathbb{M}$, $\mathbb{F}$, and $\mathbb{T}$ denote the index sets of the microphones, frequency bins, and time frames, respectively.

3.1. Feature set

With VarArray, we use a variable-size feature set, $\{z_{mt}\}_{m\in\mathbb{M}}$, for each time frame t instead of stacking all the features obtained from different microphones in a particular order. This prevents the input features from being tied to a specific microphone array with a predefined channel indexing system. The simplest way is to use the STFT coefficients of the mth microphone as z_{mt}. Below, we propose a normalized IPD-based feature set, which outperformed the STFT-based method in our preliminary experiment.

With the proposed feature set, z_{mt} is defined as the concatenation of the magnitude of an average spectrum and the channel-wise IPD with respect to the average spectrum. That is, we have $z_{mt} = [z_{mt}^{\text{mag}}, z_{mt}^{\text{ipd}}]$, where $z_{mt}^{\text{mag}} = (|\bar{X}_{ft}|^2)_{f\in\mathbb{F}}$ and $z_{mt}^{\text{ipd}} = \angle(X_{mft}/\bar{X}_{ft})_{f\in\mathbb{F}}$ with $\bar{X}_{ft}$ being the average of the STFT coefficients over the microphones, $\{X_{mft}\}_{m\in\mathbb{M}}$. Both the magnitude and IPD features are normalized at the window level to reduce random fluctuations, whose effectiveness was experimentally shown in [8].

3.2. Speech separation model

Now, we turn our attention to the modeling side of this work. The objective is to construct a speech separation model that receives the feature sequence $(z_{mt})_{t\in\mathbb{T}}$ from all the input channels and generates TF mask M_{sft} over the input segment, where s denotes the source index and takes a value in $\{0, \cdots, 3\}$. The first two sources (i.e., $s \in \{0, 1\}$) correspond to the two speakers. The other two sources are used to capture the stationary ($s = 2$) and transient ($s = 3$) noise. This is inspired by the SSN architecture proposed in [5], where SSN stands for “speech, speech, and noise”. It is required that the output be invariant to the input channel permutation and that the model be applicable to any M value.

Several previous studies applied a split-apply-combine (SAC) approach. That is, the same speech separation model is applied to each input channel independently. Then, the TF masks obtained from all the M channels are combined, for example, by taking their average [10, 11]. The output mask order may be realigned across the M channels before the average computation [24]. This approach has two drawbacks. First, it does not take advantage of the nonlinear correlation that may exist between the features of different channels. Second, applying the same model to each channel intensifies the computational cost as the number of microphones increases.

VarArray was designed to overcome the shortcomings of the SAC approach. Figure 1 shows the block diagram of the model. As with SAC, a conformer block is applied to each channel independently at each layer. However, the VarArray model has two key differences to address the two problems mentioned above.

First, TAC layers are interspersed between the conformer blocks. TAC aggregates and fuses the features of all input channels in a non-linearly transformed space and feed the outcome back to the individual channels to combine the information across the channels efficiently in a permutation-invariant fashion. We use a slightly

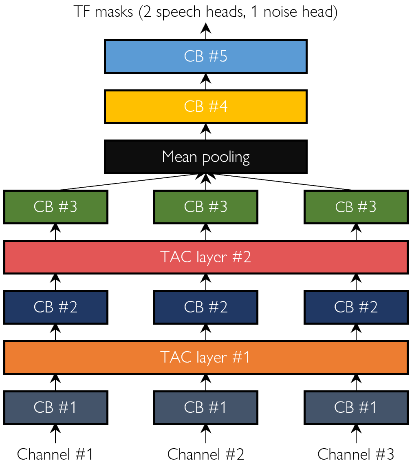

Fig. 1. Block diagram of VarArray model. Blocks with the same color share parameters. CB: conformer block.

different version of TAC than the original proposal [14]. Specifically, denoting the input sequence of the mth channel by $(\boldsymbol{o}_{mt}^{\text{in}})_{t\in\mathbb{T}}$, we compute the output for this channel, $(\boldsymbol{o}_{mt}^{\text{out}})_{t\in\mathbb{T}}$, as follows:

$$\boldsymbol{o}_{mt}^{\text{out}} = \left[\text{ReLU}\big(A\boldsymbol{o}_{mt}^{\text{in}}\big), \frac{1}{M}\sum_{\mu\in\mathbb{M}} \text{ReLU}\big(B\boldsymbol{o}_{\mu t}^{\text{in}}\big)\right], \qquad (1)$$

where matrices A and B are linear transforms.

Secondly, instead of maintaining the M channels all the way to the last layer, we merge them at an earlier layer. In the example shown in Fig. 1, the model has only one stream after the third conformer block. This significantly reduces the network computation cost.

3.3. End-to-end optimization

The VarArray model can be further optimized with respect to an ASR-based loss function. This is possible with the multi-input multi-output architecture of [25], which proposed to combine the speech separation TF mask estimator, MVDR beamformer, feature transformation, and sequence-to-sequence (S2S) ASR model to create an E2E neural network model and perform back-propagation all the way down to the speech separation model. We start the E2E optimization with a VarArray model pretrained with simulated multi-channel data. During training, we minimize the uPIT-style cross-entropy loss between two predicted hypotheses and reference transcriptions, with the back-end ASR model parameters frozen. This allows real transcribed meetings to be used to fine-tune the model parameters and mitigate the mismatch between the simulation and test conditions.

To make the E2E model equivalent to the CSS scheme, we add the sliding window stitching process (see Section 2.1), TF mask sparsification, and gain adjustment (see Section 2.2 for both) to the model. The forward pass processes each multi-channel training sample as follows. The training sample is first fed to the VarArray speech separation model to generate the four TF masks described in Section 3.2, and they are sparsified by retaining only the most dominant mask. Then, two speech separated signals are computed by using the original multi-channel signal and the MVDR filters derived from the sparse TF masks. The generated signals are then chunked into overlapping windows, and the gain adjustment is applied to each windowed signal to equalize the beamformed signal magnitudes with those of the signals that would be obtained by directly applying the TF masks to the input signal. The windowed beamformer signals are stitched together to form the two-channel separated signals that are fed to an ASR model. We used the attention-based encoder-decoder model of [26] in our experiments. Note that these additional elements, especially the gain adjustment, are critical as our training set contains many single-talker samples.

4. EXPERIMENTAL RESULTS

We conducted experiments to measure the impact of the proposed array-geometry-agnostic speech separation model on the word error rate (WER) of a meeting transcription system. Unlike conventional practices of using signal-based metrics or the WER for pre-segmented speech mixtures, this allowed us to capture the practical usefulness of the speech separation model. Our meeting transcription system, the evaluation tasks, and the results are described below.

4.1. System

We employed the audio-visual meeting transcription system of [6]. The vision and speaker diarization modules of the original system were removed to focus on the ASR accuracy. The processing flow was as follows. First, multi-input multi-output dereverberation was performed in real time with a weighted linear prediction method [27]. Then, CSS was applied by using the proposed speech separation model to generate two output signals that were deemed overlap-free. A 1.6-second window was used with a shift of 0.4 seconds, i.e., the stitching was performed by using each 1.2-second overlapping segment. Finally, ASR was applied to these signals independently, and the output transcriptions were merged. Microsoft's internal system was used for ASR.

The speech separation training set consisted of 438 (219 × 2) hours of speech mixtures that were synthesized based on clean speech signals, noise samples, and room impulse responses (RIRs). The simulation was carried out by following the recipe of [6] to create a balanced mix of the single- and two-speaker training samples as well as the four overlap patterns described in [8]. The source speech signals were taken from WSJ SI-284. The RIRs were generated by the image method [28]. Both isotropic and directional noise signals were added to the reverberant speech. The isotropic noise samples were synthesized with the algorithm of [29] while the directional noise data were taken from MUSAN [30]. The RIRs and the isotropic noise samples were generated for two microphone array geometries, namely AMI and MS (see Section 4.2). For each geometry, 219 hours of audio were generated (which is why the size of the first training set was 438 hours).

The VarArray model was configured as follows. As shown in Fig. 1, our model consisted of conformer blocks and TAC layers. Each conformer block comprised five conformer layers, each with four attention heads, 64 dimensions, and 33 convolution kernels. There were a total of five conformer blocks, three of which were applied to each stream in parallel. For TAC, matrices A and B of Eq. (1) halved the number of dimensions for the TAC layer to preserve the dimensionality. During training, the number of channels (and the microphones to be used) were randomly picked between 3 and 7 for each mini-batch to expose the model to a variety of array geometries and inter-microphone spacing patterns. The training was performed with a mini-batch of 48 four-second sequences.

Table 1. %WERs for AMI-dev and MS. Geometry-dependent models (labeled as Fixed) were trained only for full-array setting of each data set. Training settings were optimized for MS.

Data	#Mics	Baselines			VarArray
		BF1	SAC	Fixed	
AMI-dev	8	25.0	18.3	18.8	**17.7**
	4	25.5	18.8	—	**18.5**
MS	7	17.6	16.0	16.0	**15.5**
	3	17.8	16.9	—	**15.8**

4.2. Tasks

The evaluation was carried out by using two multi-microphone meeting transcription tasks: AMI (or, more precisely, AMI-MDM) and an internal meetings collection, dubbed as MS. For AMI, the "full-corpus ASR" partition [31] was employed. The audio data were recorded with an eight-channel circular microphone array with a radius of 10 cm. Eight-channel and four-channel conditions were considered, where the latter used only the microphones with odd number indices. MS is an extension of the test set used in [6]. The same seven-channel microphone array was used. Six microphones were arranged in a circle with a radius of 4.25 cm, and the last microphone was located at the center. Seven-channel and three-channel conditions were considered, where the latter used the three microphones including the center microphone and constituting a small equilateral triangle. A total of 60 sessions were recorded (150K words in the reference transcriptions). The WER was estimated with asclite for AMI while our internal scoring tool was used for the MS test set. We used MS to tune system's hyper-parameter values. The AMI development set was actually used as an "evaluation set".

4.3. Results

Table 1 shows the WERs of the proposed speech separation method and three baseline systems. The first baseline system (BF1) performed super-directive single-output beamforming with real-time beam-steering using our internal beamformer, followed by ASR and thus could not handle speech overlaps. This beamformer performed was as effective as BeamformIt [32] in our internal test. The second baseline system was based on SAC and applied a single-stream separation model to each input stream of Fig. 1 with the magnitude and IPD pair as input to obtain TF masks from each stream. As with [24], the output speech orders were aligned across the streams, and then the TF masks were averaged to be used for beamforming. This separation model had the same number of conformer blocks as our proposed VarArray system. It was trained on the same data set as ours by randomly choosing the input stream. The third baseline system used a concatenation of the magnitude spectra of the first microphone and the IPDs between the first microphone and the rest of the microphones. The model was trained for AMI and MS separately based on the corresponding subsets of the training set due to the input dependency on the number of microphones.

For all the four conditions, our system significantly outperformed the second baseline system, showing VarArray's capability of effectively leveraging the spatial information for different microphone array geometries. The proposed model also outperformed the array-geometry-dependent models for both test sets. This could be largely attributed to the fact that the array-geometry-agnostic model was able to take advantage a larger training set.

Two additional VarArray models were further trained to examine the performance dependency on the array configuration of the training set. One model was trained on the subset that was based on the AMI geometry RIRs. The other model used the MS-geometry portion of the training set. Table 2 shows the experimental results. It was observed that the separation model trained with both microphone array configurations performed the best for both test sets. Nonetheless, the other two models also showed good generalization ability to unseen array geometries. It is noteworthy that the model trained on the AMI-geometry subset performed slightly better than the one trained on the MS-geometry subset. We presume that this is because the AMI-geometry portion of the training set had greater diversity in terms of the inter-microphone distances.

Table 2. Performance dependency on microphone array configuration of training set.

Train Test	AMI	MS	AMI+MS
AMI-dev	18.0	18.8	**17.7**
MS	15.7	15.8	**15.5**

Table 3. Impact of end-to-end optimization on %WER.

Pre-Train	E2E optim	AMI-dev		AMI-eval	
		w/o ovlp	w/ ovlp	w/o ovlp	w/ ovlp
MS		15.3	18.7	17.4	22.2
MS	✓	**14.8**	**18.2**	**16.8**	**21.5**
AMI+MS		15.5	17.8	17.0	20.5
AMI+MS	✓	**14.7**	**17.5**	**16.8**	**20.4**

Table 3 shows the E2E optimization experiment results for the AMI development and evaluation sets. We segmented the original long-form AMI training recordings by silence positions, removed the segments shorter than 10 seconds, and picked only the segments consisting of one or two speakers, which resulted in a 8.52-hour training set. We experimented two seed models: one trained on the MS-geometry data, one on the AMI- and MS-geometry data. Note that, in this experiment, we used a larger pool of clean speech than in the previous experiments, including samples from our internal data, while we used the same RIR and noise data. The training set size for the seed models was 1,500 hours for the MS- and AMI-geometry portions each. Therefore, the WER numbers are slightly different from those of the previous experiments. Consistent WER gains were observed, suggesting the E2E optimization using the small amount of real transcribed data to help mitigate the mismatch between the training and test conditions. Larger gains were observed for the model pre-trained on the MS-geometry data, indicating that the E2E optimization addressed both the geometry difference and the mismatch between the simulated and real data. It is also noteworthy that the gains were more prominent for non-overlap regions. This could mean that the E2E fine-tuning made the separated signals more friendly to ASR rather than removing interfering signals.

5. CONCLUSION

We described VarArray, an array-geometry-agnostic speech separation model. Extensive evaluation was conducted with a CSS framework in two real meeting transcription tasks. The proposed model outperformed the SAC approach by efficiently leveraging the spatial information. It was also shown that the VarArray model could generalize to unseen array geometries well and that the E2E optimization could mitigate the mismatch between the training and test conditions. The array-geometry-agnostic modeling is useful for production. In parallel to this work, we examined its impact on personalized noise reduction [33]. Further investigation in different tasks is desired.

6. REFERENCES

[1] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," *arXiv:2010.10504 [eess.AS]*, 2020.

[2] Z. Tüske, G. Saon, and B. Kingsbury, "On the limit of English conversational speech recognition," in *Proc. Interspeech*, 2021, pp. 2062–2066.

[3] P. Szymański, P. Żelasko, M. Morzy, A. Szymczak, M. Żyła Hoppe, J. Banaszczak, L. Augustyniak, J. Mizgajski, and Y. Carmiel, "WER we are and WER we think we are," *arXiv:2010.03432 [cs.CL]*, 2020.

[4] M. Del Rio, N. Delworth, R. Westerman, M. Huang, N. Bhandari, J. Palakapilly, Q. McNamara, J. Dong, P. Żelasko, and Miguel Jetté, "Earnings-21: a practical benchmark for ASR in the wild," in *Proc. Interspeech*, 2021, pp. 3465–3469.

[5] T. Yoshioka, H. Erdogan, Z. Chen, X. Xiao, and F. Alleva, "Recognizing overlapped speech in meetings: a multichannel separation approach using neural networks," in *Proc. Interspeech*, 2018, pp. 3038–3042.

[6] Takuya Yoshioka, Igor Abramovski, Cem Aksoylar, Zhuo Chen, Moshe David, Dimitrios Dimitriadis, Yifan Gong, Ilya Gurvich, Xuedong Huang, Yan Huang, Aviv Hurvitz, Li Jiang, Sharon Koubi, Eyal Krupka, Ido Leichter, Changliang Liu, Partha Parthasarathy, Alon Vinnikov, Lingfeng Wu, Xiong Xiao, Wayne Xiong, Huaming Wang, Zhenghao Wang, Jun Zhang, Yong Zhao, and Tianyan Zhou, "Advances in online audio-visual meeting transcription," in *Proc. Worksh. Automat. Speech Recognition, Understanding*, 2019, pp. 276–283.

[7] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: Dataset and analysis," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7284–7288.

[8] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5739–5743.

[9] Z.-Q. Wang, P. Wang, and D. Wang, "Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2001–2014, 2021.

[10] X. Chang, W. Zhang, Y. Qian, J. Le Roux, and S. Watanabe, "End-to-end multi-speaker speech recognition with transformer," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6134–6138.

[11] W. Zhang, C. Boeddeker, S. Watanabe, T. Nakatani, M. Delcroix, K. Kinoshita, T. Ochiai, N. Kamo, R. Haeb-Umbach, and Y. Qian, "End-to-end dereverberation, beamforming, and speech recognition with improved numerical stability and advanced frontend," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6898–6902.

[12] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[13] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, and M. Zhou, "Continuous speech separation with conformer," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 5749–5753.

[14] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6394–6398.

[15] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction*, S. Renals and S. Bengio, Eds., 2006, pp. 28–39.

[16] Y. Yemini, E. Fetaya, H. Maron, and S. Gannot, "Scene-agnostic multi-microphone speech dereverberation," in *Proc. Interspeech*, 2021, pp. 1129–1133.

[17] Siyuan Zhang and Xiaofei Li, "Microphone array generalization for multichannel narrowband deep speech enhancement," in *Proc. Interspeech*, 2021, pp. 666–670.

[18] D. Wang, Z. Chen, and T. Yoshioka, "Neural speech separation using spatially distributed microphones," in *Proc. Interspeech*, 2020, pp. 339–343.

[19] T. Yoshioka, Z. Chen, C. Liu, X. Xiao, H. Erdogan, and D. Dimitrios, "Low-latency speaker-independent continuous speech separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6980–6984.

[20] T. von Neumann, K. Kinoshita, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach, "Graph-PIT: Generalized permutation invariant training for continuous separation of arbitrary numbers of speakers," in *Proc. Interspeech*, 2021, pp. 3490–3494.

[21] O. Çetin and E. Shriberg, "Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: Insights for automatic speech recognition," in *Proc. Interspeech*, 2006, pp. 293–296.

[22] 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, Language Process.*, vol. 25, no. 10, pp. 1901–1913, 2017.

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

[24] Z.-Q. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 27, no. 2, pp. 457–468, 2019.

[25] X. Chang, W. Zhang, Y. Qian, J. Le Roux, and S. Watanabe, "MIMO-Speech: End-to-end multi-channel multi-speaker speech recognition," in *Proc. Worksh. Automat. Speech Recognition, Understanding*, 2019, pp. 237–244.

[26] X. Wang, N. Kanda, Y. Gaur, Z. Chen, Z. Meng, and T. Yoshioka, "Exploring end-to-end multi-channel ASR with bias information for meeting transcription," in *Proc. Spoken Language Process. Worksh.*, 2021, pp. 833–840.

[27] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 10, pp. 2707–2720, 2012.

[28] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[29] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *J. Acoust. Soc. Am.*, vol. 122, no. 6, pp. 3464–3470, 2007.

[30] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: a music, speech, and noise corpus," *arXiv:1510.08484 [cs.CD]*, 2015.

[31] https://groups.inf.ed.ac.uk/ami/corpus/datasets.shtml.

[32] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 7, pp. 2011–2022, 2007.

[33] H. Taherian, S. E. Eskimez, T. Yoshioka, H. Huang, Z. Chen, and X. Huang, "One model to enhance them all: array geometry agnostic multi-channel personalized speech enhancement," *arXiv:2110.10330 [eess.AS]*, 2022.