Interspeech 2024
1-5 September 2024, Kos, Greece

Enhanced Deep Speech Separation in Clustered Ad Hoc Distributed Microphone Environments

Jihyun Kim[*1], *Stijn Kindt*[*2], *Nilesh Madhu*[2], *Hong-Goo Kang*[1]

[1]Department of Electrical and Electronic Engineering, Yonsei University, South Korea
[2]IDLab, Ghent University - imec, Ghent, Belgium

jihyun93815@dsp.yonsei.ac.kr, stijn.kindt@ugent.be, nilesh.madhu@ugent.be, hgkang@yonsei.ac.kr

Abstract

Ad-hoc distributed microphone environments, where microphone locations and numbers are unpredictable, present a challenge to traditional deep learning models, which typically require fixed architectures. To tailor deep learning models to accommodate arbitrary array configurations, the Transform-Average-Concatenate (TAC) layer was previously introduced. In this work, we integrate TAC layers with dual-path transformers for speech separation from two simultaneous talkers in realistic settings. However, the distributed nature makes it hard to fuse information across microphones efficiently. Therefore, we explore the efficacy of blindly clustering microphones around sources of interest prior to enhancement. Experimental results show that this deep cluster-informed approach significantly improves the system's capacity to cope with the inherent variability observed in ad-hoc distributed microphone environments.

Index Terms: Ad-hoc microphones, Distributed microphones, Acoustic sensor networks (ASN), Multi-channel speech processing, Speech Separation, Time-domain approach

1. Introduction

In acoustic sensor networks (ASNs), multiple microphones can be arbitrarily distributed throughout a given space. This distributed configuration provides extensive spatial coverage in contrast to compact microphone arrays [1]. ASNs are also becoming more common in daily life, with the growing number of devices equipped with one or multiple microphones, like smartwatches, smartphones, laptops and smart glasses. While speech processing [2–4], speaker localization [5, 6] and speaker verification [7] have made advancements utilizing this extra spatial information, there remains much to be explored in this field to *fully* capitalize on the possibilities it opens up.

ASNs, particularly those deployed in an ad-hoc manner, have extra challenges associated with them. Firstly, the number of microphones and their respective positions are not known and may vary throughout its operation due to environmental changes. These changes can include devices entering or leaving the environment, or moving within the space. Secondly, due to the potentially widely-distributed nature of the microphones, the same speech signal can be captured at two different microphones at very different time instances. Also, the microphones operated on independent clocks, leading to discrepancies in sample rate offsets (SROs) and sample time offsets (STOs). Lastly, all the microphones could have very different characteristics, i.e. frequency response and directivity. While the latter two can be robustly solved by other methods [8, 9], the core challenge of speaker separation endures.

One solution previously proposed by Gergen *et al.* [10, 11] is to first cluster the microphones either around the speakers or into a background (noise) cluster. Subsequently, cluster information is leveraged within classical signal enhancement frameworks, demonstrating superiority over methods like optimal microphone selection. The usefulness of clustering has also been shown in [12], where incorporating microphones that are far away from a target speaker (from outside the cluster) can degrade the result. To date, cluster-based separation techniques have only been investigated within the realm of classical signal processing. However, we hypothesize that deep neural network-based separation methods could also benefit from the clustering, potentially surpassing the performance of classical methods.

The Transform-Average-Concatenate (TAC) layer [13] was previously proposed for deep, array-agnostic, *compact* array processing. This layer is introduced between blocks that individually process the inputs for each microphone channel and, consequently, acts as the information sharing layer between microphone channels in a permutation and number invariant manner. Thus, it can handle the challenge of unknown array geometries. TAC has been successfully combined different architectures, e.g., dual-path recurrent neural network (DPRNN) [14] and VarArray [15], where Conformers [16] are used for time-frequency processing. Alternative array-agnostic methods, like [17], use multi-head cross-attention to share information across microphones. However, only limited research [17, 18] has been conducted specifically on distributed microphone setups. Their investigation revealed both the potential benefits and the intricate challenges associated with employing variably located microphones for speech processing. A significant gap identified is the need for better methods to utilize the spatial diversity of the microphones in a more informed manner.

This paper proposes a novel approach that incorporates the blind microphone clustering techniques into *cluster-informed*, array-agnostic deep learning methodologies. In short:

1. Initially, in the ad-hoc distributed microphone environment, a blind spatial-statistics-based clustering approach [19] is employed to cluster microphones around the active speakers. Additionally, the clustering also estimates a *pseudo* reference microphone, where the target speech should be the most dominant in all microphones of that cluster.
2. Then a deep learning-based network exploits spatial information from all microphones within each speaker-dominated cluster to extract the underlying target speech. As we demonstrate, the optimal configuration exploits, in addition to spatial information from all microphones within the cluster, the benefit offered by selecting a robust reference microphone.

This work is supported by the Research Foundation - Flanders (FWO) under grant number G081420N

[*] Equal contributions, shared first author

10.21437/Interspeech.2024-1378

Additionally, we propose a training data generation method to simulate clustered data without actually executing the clustering. This is important to save considerable training time and ensures that the training is independent of the specific clustering algorithm employed. The deep separation method will be compared to the classical processing methods and ablation studies will show the effectiveness of the proposed method compared to alternative deep learning structures that do not use the cluster information to its fullest potential. The paper will first overview the classical techniques, where a brief explanation of the clustering and separation method is given in Sec. 2. Then the proposed deep architecture is explained in Sec. 3, and evaluated in Sec. 4. Sec. 5 concludes the paper.

2. Classical Methods

2.1. Clustering

Ensuring robust clustering is essential for distributed microphone techniques [20, 21]. Based on the findings of the comparative study presented in [22], the full bandwidth, coherence-based clustering method [19] is chosen. This algorithm, represented by the first two steps in Fig. 1, uses the pairwise magnitude squared coherence between all M microphones as features, denoted by $\mathcal{F}$, and organizes them into a matrix $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{M\times M}$. Then, non-negative matrix factorization (NMF) [23] is utilized to cluster the microphones by decomposing this matrix as: $\boldsymbol{\mathcal{C}} = \boldsymbol{B}\boldsymbol{B}^T \odot (\mathbf{1} - \boldsymbol{I}) + \boldsymbol{I}$, where $\odot$ is the element-wise (Hadamard) product, $\boldsymbol{I}$ denotes the identity matrix, $\mathbf{1}$ is the all-ones matrix and $\boldsymbol{B} \in \mathbb{R}^{M\times C}$ is the cluster matrix. This matrix contains all fuzzy membership values (FMVs) of each microphone towards each cluster, where $\boldsymbol{B}_{mc}$ represents the contribution of microphone m to cluster c. For hard clustering, microphones are then attributed to the cluster where their contribution is highest. Additionally, for each cluster, a reference microphone is identified as the microphone with the highest fuzzy membership value for that cluster. The number of clusters C is one greater than the number of speakers, where the last cluster collects microphones mostly dominated by noise and reverberations. Both the (hard) cluster and the reference microphone will prove invaluable for *informed* speech separation.

2.2. Separation

In Fig. 1, the relation between the different classical cluster based separation methods are shown: (1) initial masks are estimated by comparing the amplitude of the short-time Fourier transform (STFT) bins across all reference microphones within each cluster, exploiting sparsity and disjointness of speech [24]. (2) Then, relative time delays $\widehat{\tau}_{m,c}$ are estimated on the masked cluster signals, and used for delay and sum beamforming (DSB) of all clustered microphone signals. (3) A microphone weighting based on the fuzzy values can be included in the fuzzy membership value aware DSB (FMVA-DSB). (4) Lastly, a postfilter can obtained by comparing the beamformed signal of each cluster. For a detailed overview, we refer to [10, 11]. All the methods succeed in a better foregrounding of target speakers, but the resulting audio quality is poor. The masks of initial and postfilter are binary, distorting the signal, and the simple beamformers cannot suffciently cancel the interferer.

3. Proposed Method

As illustrated in Fig. 2, our proposed method clusters ad-hoc distributed microphones around speakers and selects a reference microphone for each cluster. This is done as described

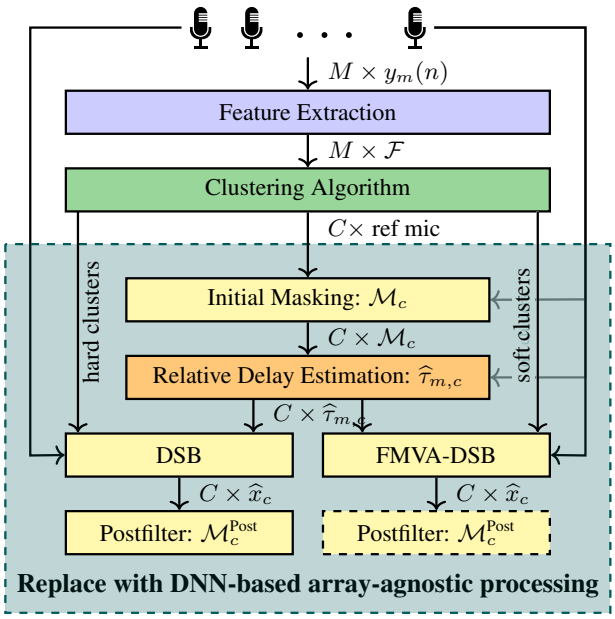

Figure 1: *Scheme of the classical cluster-based source separation method. The grayed region is replaced by our proposed separation.*

in Sec. 2.1. Microphones of each cluster are then processed through a deep learning-based separation network, which consists of Encoder, Separator and Decoder.

By clustering first, each cluster can be independently processed - removing the need for separation techniques such as permutation invariant training (PIT) [25]. We only need to extract the *dominant* source of each cluster with a multi-channel time-domain network. The network architecture is based on the VarArray structure [15], where Conformers are swapped with dual-path transformer networks (DPTNets) [26]. The use of the DPTNet allows for a computationally efficient method to process both local and global information, leading to a comprehensive and accurate representation of the acoustic scene.

Encoder. The Encoder transforms raw multi-channel speech signals $x \in \mathbb{R}^{M\times 1\times T}$ into a high-dimensional feature $h \in \mathbb{R}^{M\times N\times T}$ using a 1-D convolution layer. Additionally, the Encoder incorporates a segmentation strategy derived from DPTNet, allowing for more precise handling of both local and global dependencies within the speech signal. It splits the hidden feature h into overlapped chunks of length K with a hop size of $K/2$. The hidden feature h is thus a 4-D tensor $h_0 \in \mathbb{R}^{M\times N\times K\times P}$, where P is the number of chunks. This design choice ensures that the network both preserves the detailed temporal structure of the speech signal and increases its receptive field. This is essential to compensate for the relatively long time delays in widely distributed microphone settings.

Separator. The separator combines TAC layers and DPTNets within its processing chain for spatial and temporal processing respectively. Similar to the structure of VarArray, 3 DPTNets are interleaved with 2 TAC layers, followed by a mean pooling and 2 DPTNets. The first DPTNets process each microphone individually, where the TAC layers combine the microphone information. Mean pooling reduces computational complexity for the following (single channel) DPTNets. The separator produces a mask, which is applied to an encoder embedding. Here the suitability of the clustering, more specifically the indication

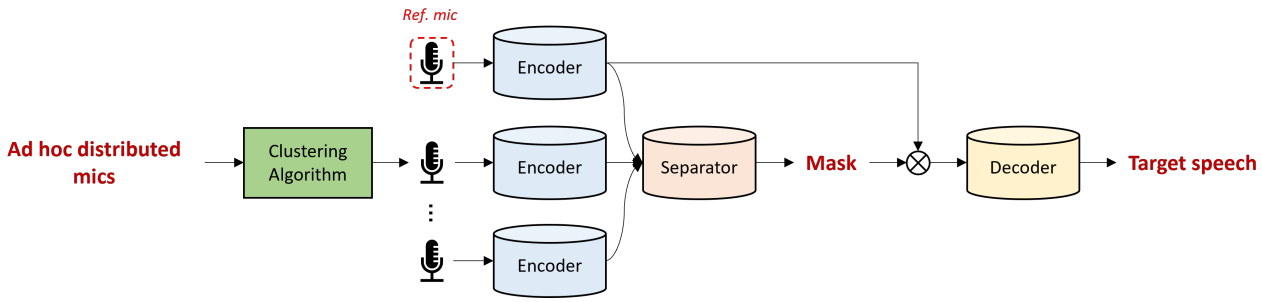

Figure 2: *Overall system architecture of the proposed deep, array-agnostic, target extraction approach. The reference microphone is identified from the clustering. For separation, the mask is applied to the embedding of the reference microphone.*

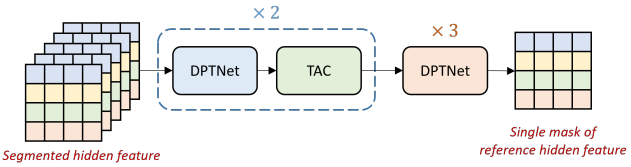

Figure 3: *Detail structure of the separator. For multi-microphone processing, a single module operates in parallel across the microphones.*

of a reference microphone, is again clear. If this information is present, the hidden feature of the reference microphone can be selected for masking.

We propose further changing the mean pooling layer by selecting the embedding from the reference microphone. Firstly, this embedding should be the most dominated by the target speech. Secondly, it reduces the computational cost since no DPTNets operations are applied on the non reference embeddings. The proposed method is depicted in Fig. 3. We will term this proposed style, while the previously described networks VarArray style in the ablation studies of Sec. 4.3.

Decoder. The Decoder reconstructs the separated speech signals from the enhanced high-dimensional latent representation. In a process that mirrors the encoder, the decoder employs overlap and add, followed by a 1-D transposed convolution layer, to transform the enhanced hidden feature back into the time-domain yielding clear, distinct speech tracks.

4. Experimental Evaluation

In this section, we want to show that deep networks improve upon classical separation methods and that cluster information is essential for good separation. Therefore, next to comparison with the classical methods, we include ablation studies where all microphones are used as input to the neural network (*unclustered version*), an ablation study where only the reference microphone (*single microphone*) is used as input, and study the difference between VarArray style and the proposed style.

The unclustered version is trained with PIT loss to separate the different speakers. Since no reference mic is known, a random microphone is selected of which the encoding embeddings are taken for the decoder. Initial experiments showed that averaging over the embeddings performs worse. This highlights why clustering in widely distributed microphones is beneficial.

The ablation study with the single microphone method, where no TAC layers are needed, indicates whether the spatial diversity, provided by the multiple microphones within a cluster, is valuable. The VarArray style ablation study reveals whether incorporating the reference microphone information within the network structure improves the final result.

We will begin by outlining the datasets used for training and evaluation, as well as the specific parameter choices, before delving into the results.

4.1. Dataset

To train our network, we utilized the WSJ0-2mix [27] clean speech dataset, and convolved them with the shoebox room impulse responses (RIRs), generated by the image source model of gpuRIR [28]. White noise was added at SNRs uniformly sampled between 0 dB and 20 dB. A variety of different rooms (dimensions and reverberation times) are generated to promote generalizability, totaling 10,080 different scenarios. For this work, the simulation was limited to cases where the two speakers were located in different halves of the room. A total of 16 microphones were simulated for each scenario, where for each source there are at least 3 microphones within its critical distance, consistent with previous cluster based separation work [21]. We iterate over all RIRs for each epoch and select a random clean speech sample on the fly to increase diversity. Also, for each batch, a random selection of microphones – between 8 and 16 microphones – is chosen, to expose the network to different numbers of microphones and increase the total number of possible scenarios. However, it is ensured that the microphones within the critical distance are kept during selection.

Clustered Training Dataset. However, clustering on the fly would waste valuable training time since NMF is an iterative method. Also, this might make the network dependent on the clustering algorithm. To alleviate these problems, a second dataset of clustered RIRs is generated, where microphone positions are simulated as if they could have originated from clustering. Two sources are still sampled in different room halves, but the microphones are no longer sampled in the whole room. Three microphones are still placed within the critical distance of the speaker, and one is selected as the reference microphone. 4 other microphones are simulated in a $2m \times 2m$ square centered around the speaker. During training, a random number of microphones between 3 and 7 is chosen for generalizability to unknown microphone numbers. The unclustered dataset is utilized to train the unclustered version, while the clustered dataset is used to train all other networks.

Evaluation Dataset. To assess the real-world applicability of the model, the realistic SINS dataset [29], simulated with a CATT model, is used. The evaluation set is done similarly to [21,22].Two speaker positions are selected in opposite halves of the room and 16 microphones are distributed over the room. For each source, at least 3 microphones are within the critical distance. If the speakers are both sampled towards the middle of

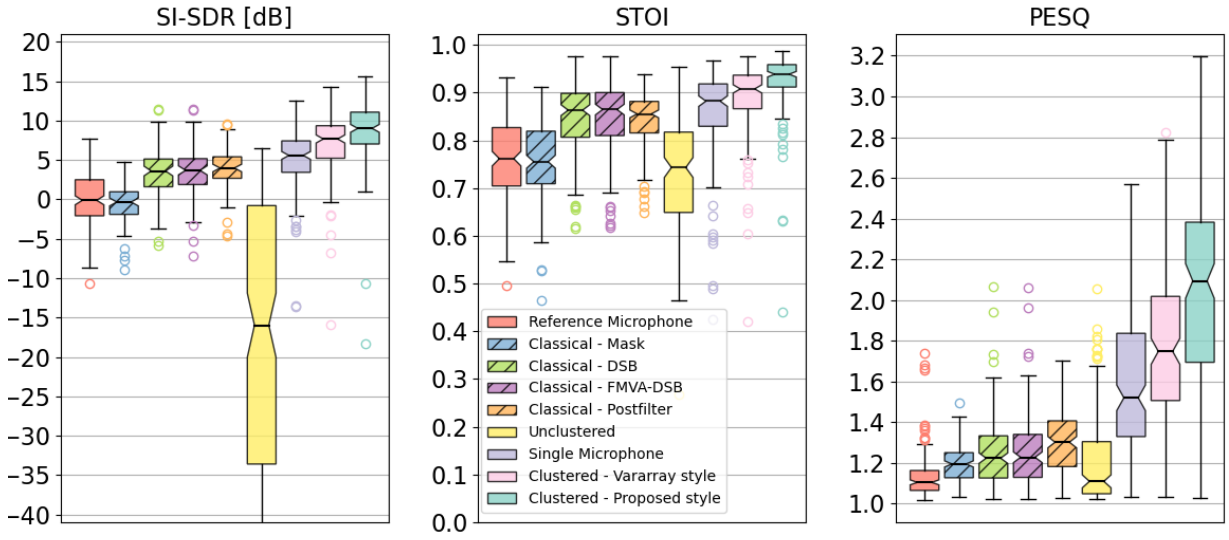

Figure 4: *SI-SDR, STOI and PESQ for the different separation methods (higher is better)*

the room, they *could* still be very closely spaced. Dry speech is taken from the LibriSpeech dataset [30]. White noise is added, at an SNR of 10 dB with respect to the middle of the room. The actual SNR at individual microphones can differ greatly.

Good performance on real clustered data would validate the clustered training dataset. Also, since this dataset differs significantly in realism and sound sources, it can demonstrate the model's performance and generalization capability for environments that closely mimic actual speech separation challenges.

4.2. Experiment Setup

For the Encoder and Decoder, we selected a kernel size of 8 samples with a stride of 50%. We use a segment size K of 250 on the segmentation for dual-path processing. We set the feature dimension of the separation network to be 64. We use 4 attention heads on each transformer layer in the DPTNet. The training criterion we used is the Scale Invariant Signal to Distortion Ratio (SI-SDR) [31] loss We used Adam optimizer and the training process began with an initial learning rate of 0.125, with a strategy to halve the learning rate if the validation loss doesn't decrease for three epochs. The total number of parameters in our proposed model is 2.23 M.

4.3. Experiment Results

To assess the performance, we employed three objective metrics: Scale Invariant Signal to Distortion Ratio (SI-SDR) [31], Perceptual Evaluation of Speech Quality (PESQ) [32], and Short-Time Objective Intelligibility (STOI) [33]

Fig. 4 shows the results of the different separation methods. The *reference point* is given by the metrics computed on the *unprocessed* reference microphone signals. Firstly, it is clear, from the SI-SDR and PESQ, that combining information across *all* microphones (unclustered version) does not perform well. This indicates that the general, uninformed nature of TAC – which combines features from all microphone signals similarly, independent of the underlying speaker or background noise dominance at that microphone – does not suit distributed scenarios. Additionally, the lack of a good selection mechanism for hidden features on which the mask is applied, makes it hard for the network to generalize to other situations: the SI-SDR performance is very poor even though it was trained to maximize this metric.

Picking the unprocessed microphone closest to each speaker – the reference microphone in the clustering algorithm – outperforms the unclustered deep leaning method and shows decent intelligibility (STOI). However, the output can be further improved. The classical methods do indeed increase the performance on all metrics, except for the initial masking – which is anyway mainly used for a robust, relative time delay estimation.

Using cluster informed deep learning algorithms significantly outperforms the classical methods. Using the reference microphone as input for a single channel model, where only DPTNets are sequentially applied, gives a big performance boost, most notably the big increase in PESQ. However, the method does not exploit the spatial diversity provided by the clustered microphones. When considering the inputs from clustered microphones, the metrics show that, unlike the unclustered version, TAC is effective. All microphones are dominated by the same source, removing target ambiguity. The performance of the clustered methods also supports the validity of the proposed data generation scheme.

The results also show that it is worthwhile to let the network prioritize the embeddings from the reference microphone. VarArray style averages the features over all the microphones before continuing with the single channel portion of the network. The proposed style takes the reference microphone as input for the single-channel portion of the network. This simple information inclusion in the design increases the performance significantly on all three metrics.

Specific scenarios, where the clusters are plotted and audio of the different separation techniques is present, can be found at https://aspire.ugent.be/demos/INTERSPEECH2024SK/ .

5. Conclusion

In this paper, we introduced a novel approach for speech separation in ad hoc distributed microphone environments, combining coherence-based clustering methods with deep learning networks. Our experiments on realistically simulated RIRs show that it is essential to include cluster information in deep learning separation networks. More so, also including the reference microphone – a byproduct of the clustering method – further enhances the method. Conversely, the deep learning based separation gives a significant boost to the separation compared to classical methods. This highlights the benefits of combining traditional signal processing techniques with modern deep learning for speech processing tasks in real-world scenarios. Additionally, an efficient data generation paradigm to simulate clustered data was proposed for training such frameworks.

Future work could further increase the information the networks get from the clustering, by incorporating cross cluster information within the design of the network.

6. References

[1] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *2011 18th IEEE symposium on communications and vehicular technology in the Benelux (SCVT)*. IEEE, 2011, pp. 1–6.

[2] M. Souden, K. Kinoshita, M. Delcroix, and T. Nakatani, "Location feature integration for clustering-based speech separation in distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 354–367, 2013.

[3] D. Kim, S.-W. Chung, H. Han, Y. Ji, and H.-G. Kang, "HD-DEMUCS: General Speech Restoration with Heterogeneous Decoders," in *Proc. INTERSPEECH 2023*, 2023, pp. 3829–3833.

[4] J. Kim and H.-G. Kang, "Contrastive Learning based Deep Latent Masking for Music Source Separation," in *Proc. INTERSPEECH 2023*, 2023, pp. 3709–3713.

[5] S. Kindt, A. Bohlender, and N. Madhu, "2d acoustic source localisation using decentralised deep neural networks on distributed microphone arrays," in *Speech Communication; 14th ITG Conference*. VDE, 2021, pp. 1–5.

[6] H. Han and N. Kumar, "A cross-talk robust multichannel vad model for multiparty agent interactions trained using synthetic rerecordings," in *2024 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2024.

[7] D. Cai and M. Li, "Embedding aggregation for far-field speaker verification with distributed microphone arrays," in *2021 IEEE spoken language technology workshop (SLT)*. IEEE, 2021, pp. 308–315.

[8] T. Gburrek, J. Schmalenstroeer, and R. Haeb-Umbach, "On synchronization of wireless acoustic sensor networks in the presence of time-varying sampling rate offsets and speaker changes," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 916–920.

[9] A. Chinaev, N. Knaepper, and G. Enzner, "Long-term synchronization of wireless acoustic sensor networks with nonpersistent acoustic activity using coherence state," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] S. Gergen, R. Martin, and N. Madhu, "Source separation by feature-based clustering of microphones in ad hoc arrays," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 530–534.

[11] S. Gergen, R. Martin, and N. Madhu, "Source separation by fuzzy-membership value aware beamforming and masking in ad hoc arrays," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.

[12] I. Himawan, I. McCowan, and S. Sridharan, "Clustered blind beamforming from ad-hoc microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 661–676, 2010.

[13] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6394–6398.

[14] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[15] T. Yoshioka, X. Wang, D. Wang, M. Tang, Z. Zhu, Z. Chen, and N. Kanda, "Vararray: Array-geometry-agnostic continuous speech separation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6027–6031.

[16] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *Interspeech 2020*, 2020.

[17] D. Wang, Z. Chen, and T. Yoshioka, "Neural speech separation using spatially distributed microphones," *INTERSPEECH 2020*, pp. 2467–2471, 2020.

[18] D. Wang, T. Yoshioka, Z. Chen, X. Wang, T. Zhou, and Z. Meng, "Continuous speech separation with ad hoc microphone arrays," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 1100–1104.

[19] A. J. Muñoz-Montoro, P. Vera-Candeas, and M. G. Christensen, "A coherence-based clustering method for multichannel speech enhancement in wireless acoustic sensor networks," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 1130–1134.

[20] S. Gergen, A. Nagathil, and R. Martin, "Classification of reverberant audio signals using clustered ad hoc distributed microphones," *Signal Processing*, vol. 107, pp. 21–32, 2015.

[21] S. Kindt, J. Thienpondt, L. Becker, and N. Madhu, "Robustness of ad hoc microphone clustering using speaker embeddings: evaluation under realistic and challenging scenarios," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, p. 46, 2023.

[22] S. Kindt, M. Meeldijk, and N. Madhu, "Ad hoc distributed microphones clustering: A comparative analysis on using coherence and signal-specific features," in *Speech Communication; 15th ITG Conference*. VDE, 2023, pp. 11–15.

[23] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, 2000.

[24] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. IEEE, 2002, pp. I–529.

[25] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[26] J. Chen, Q. Mao, and D. Liu, "Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation," in *Proc. Interspeech 2020*, 2020, pp. 2642–2646.

[27] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[28] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpurir: A python library for room impulse response simulation with gpu acceleration," *Multimedia Tools and Applications*, vol. 80, pp. 5653–5671, 2021.

[29] R. Glitza, L. Becker, A. Nelus, and R. Martin, "Database of simulated room impulse responses for acoustic sensor networks deployed in complex multi-source acoustic environments." in *EUSIPCO*, 2023.

[30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[31] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[32] 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 *IEEE Intl. Conf. on acoustics, speech, and signal processing.*, vol. 2, 2001, pp. 749–752.

[33] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE Intl. Conf. on acoustics, speech and signal processing*, 2010, pp. 4214–4217.