Spatial–Temporal Decoupling for Microphone Utility Modeling in Ad-Hoc Multi-Channel Speech Separation

Junzhe Chen*, Yichi Wang*, Yahui Fu* and Tatsuya Kawahara*
* Graduate School of Informatics, Kyoto University, Kyoto, Japan
Email: {juchen, yiwang, fu, kawahara}@sap.ist.i.kyoto-u.ac.jp

***Abstract*—Ad-hoc microphone arrays enable flexible spatial deployment for multi-channel speech separation but face challenges from variable channel counts, arbitrary channel ordering, and heterogeneous microphone utility. Existing methods often assume equal channel contributions or rely on early cross-channel aggregation, letting low-utility channels degrade fused representations. We propose a Decoupled Spatial–Temporal (DeST) block for ad-hoc speech separation. Dilated convolution attention (DCA) captures channel-aware spatial features, while a squeezed temporal convolutional network (S-TCN) models intra-channel temporal features. The two paths are fused via iterative attentional feature fusion (iAFF), where spatial features reweight temporal features to achieve utility-aware integration. Experiments under diverse acoustic conditions show consistent improvements over competing methods across multiple separation metrics while maintaining competitive parameter efficiency. Visualization further validates the learned microphone utility modeling.**

***Index Terms*—Multi-channel speech separation, ad-hoc microphone arrays, microphone utility modeling**

I. Introduction

In cocktail party scenarios with multiple simultaneous speakers in noisy and reverberant environments, single-channel speech separation relies solely on time–frequency modeling and often lacks robustness [1]. Multi-channel methods [2], [3], [4], [5], [6] exploit spatial information across microphones, as speakers arriving from varied directions induce direction-dependent inter-channel differences, and have therefore attracted increasing attention.

Traditional multi-channel speech separation methods [7] typically rely on compact, fixed-geometry microphone arrays with predefined sensor positions, in which the channel count and ordering are fixed. However, these assumptions are restrictive in practice: compact geometries limit spatial resolution and degrade with increasing speaker distance, whereas fixed deployment reduces flexibility [8]. Ad-hoc microphone arrays address these limitations by eliminating the reliance on compact, fixed-geometry deployments. They comprise spatially distributed acoustic sensing nodes that can be flexibly arranged, enabling broader spatial coverage and scalable deployment [9], [10], [11]. In practice, ad-hoc arrays can be realized using commodity microphone-equipped devices (e.g., smartphones, laptops, and tablets), with captured signals transmitted to a centralized processor for joint separation.

However, ad-hoc microphone arrays introduce two key challenges: First, the number of microphones can vary, and channel ordering can be arbitrary, violating the fixed channel count and consistent ordering assumptions adopted by most multi-channel models [10], [12], [13]. Second, spatially distributed microphones experience different acoustic propagation conditions, resulting in substantial utility differences across channels. Here, microphone utility refers to the relative contribution of each microphone channel to target speech separation performance under a given acoustic condition. Such utility is affected by factors including speaker–microphone distance, relative signal-to-noise ratio (SNR), reverberation, and inter-channel redundancy [14].

To address the first challenge, transform-average-concatenate (TAC) [12] applies a shared transformation to each channel and performs global average pooling to obtain a permutation-invariant representation. However, this symmetric aggregation treats all channels equally, ignoring the substantial variability in microphone utility in ad-hoc arrays. For the second challenge, [8] models microphone utility variation via a probabilistic framework with expectation-maximization (EM)-based estimation of microphone-dependent speaker activities. However, the lack of explicit inter-channel modeling limits the exploitation of spatial information. To solve it, in [10], channel-wise self-attention is used to handle variable channel counts and arbitrary permutations. Nevertheless, performing cross-channel interaction at early stages may allow low-utility channels to influence the fused representation.

To address these limitations, we propose a Decoupled Spatial–Temporal (DeST) block for ad-hoc speech separation. The spatial path employs dilated convolution attention (DCA) to capture multi-resolution and channel-aware spatial features, while the temporal path applies a squeezed temporal convolutional network (S-TCN) [15], [16] to each channel independently to extract intra-channel temporal features. The two paths are fused via iterative attentional feature fusion (iAFF) [17], where spatial features adaptively reweight temporal features. This decoupled design avoids premature cross-channel aggregation and mitigates the influence of low-utility channels. Experiments under diverse acoustic conditions

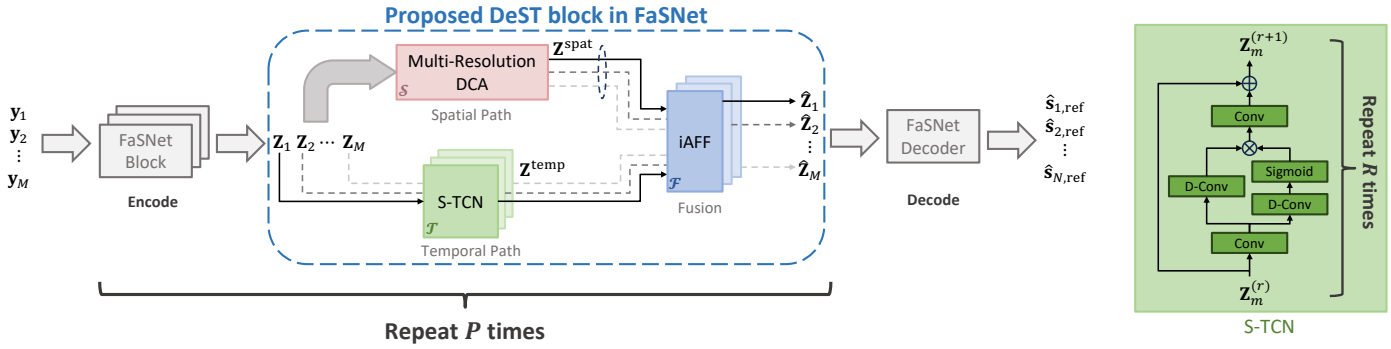


Fig. 1. Architecture of the proposed DeST block.

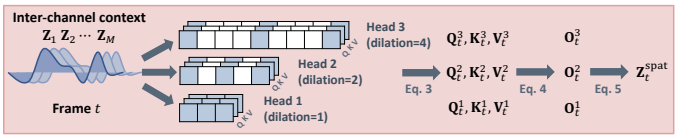


Fig. 2. Illustration of the multi-resolution feature projection mechanism in DCA.

demonstrate superior performance compared with competing methods across multiple separation metrics, while maintaining competitive parameter efficiency. Visualization further shows that the learned microphone utility weights correlate with the microphone–speaker distance.

II. Proposed Method

The overall architecture of the proposed method is shown in Fig. 1. An ad-hoc microphone array with M microphones is randomly deployed in a three-dimensional space. Let $\mathbf{s}_{i,m} \in R^L$ denote the reverberant speech signal of speaker i recorded at microphone m, where L is the number of time-domain samples, $i \in \{1,\dots,I\}$ and $m \in \{1,\dots,M\}$. The mixture signal at microphone m is defined as:

$$\mathbf{y}_m = \sum_{i=1}^{I} \mathbf{s}_{i,m} + \mathbf{n}_m, \tag{1}$$

where $\mathbf{n}_m$ represents the additive noise. The objective of this work is to estimate each speaker's signal at the reference microphone, denoted by $\{\hat{\mathbf{s}}_{i,\mathrm{ref}}\}_{i=1}^{I}$, from the observed multi-channel mixture $\mathbf{y} = \{\mathbf{y}_m\}_{m=1}^{M}$.

In this work, we adopt FaSNet [18] as the backbone separation network. The multi-channel input $\mathbf{y}$ is first encoded into an audio feature $\mathbf{Z} \in R^{M\times T\times F}$, where T denotes the number of time frames and F the feature dimension per frame. The proposed DeST blocks are integrated into the P repeated modules of FaSNet. Within each DeST block, the feature $\mathbf{Z}$ is decomposed into a spatial path and a temporal path for decoupled modeling. Specifically, the spatial operator $\mathcal{S}(\cdot)$ captures channel-aware spatial features, while the temporal operator $\mathcal{T}(\cdot)$ extracts intra-channel temporal features. The resulting features $\mathbf{Z}^{\mathrm{spat}}$ and $\mathbf{Z}^{\mathrm{temp}}$ are then fused by a fusion module $\mathcal{F}(\cdot)$ to produce the channel-aware temporal features $\hat{\mathbf{Z}}$:

$$\begin{aligned} &\textit{Spatial Path:} \quad \mathbf{Z}^{\mathrm{spat}} = \mathcal{S}(\mathbf{Z}), \\ &\textit{Temporal Path:} \quad \mathbf{Z}^{\mathrm{temp}} = \{\mathcal{T}(\mathbf{Z}_m)\}_{m=1}^{M}, \\ &\textit{Fusion Module:} \quad \hat{\mathbf{Z}} = \mathcal{F}(\mathbf{Z}^{\mathrm{temp}}, \mathbf{Z}^{\mathrm{spat}}). \end{aligned} \tag{2}$$

Finally, $\hat{\mathbf{Z}}$ is passed to the FaSNet decoder to generate the separated speech signals $\{\hat{\mathbf{s}}_{i,\mathrm{ref}}\}_{i=1}^{I}$.

A. Spatial Path

In the spatial path $\mathcal{S}$, we propose DCA to extract channel-aware spatial features $\mathbf{Z}^{\mathrm{spat}}$ from the multi-channel audio feature $\mathbf{Z}$. As shown in Fig. 2, unlike standard MHSA [10], [19], [20], which employs point-wise linear projections, DCA utilizes a bank of 1-D dilated convolutions along the intra-frame feature dimension F, projecting the input features to multi-scale Query ($\mathbf{Q}$), Key ($\mathbf{K}$), and Value ($\mathbf{V}$) tensors of dimension d_k.

For each attention head $h \in \{1,\dots,H\}$ and time frame t, the projections are defined as:

$$\mathbf{Q}_t^h, \mathbf{K}_t^h, \mathbf{V}_t^h = \text{D-Conv}^{(h)}(\mathbf{Z}_t). \tag{3}$$

The channel-wise attention for head h is then computed using a softmax over the M channels:

$$\mathbf{O}_t^h = \mathrm{Softmax}\left(\frac{\mathbf{Q}_t^h (\mathbf{K}_t^h)^\top}{\sqrt{d_k}}\right)\mathbf{V}_t^h. \tag{4}$$

To adaptively integrate the multi-resolution features from the H heads, we employ a gated fusion mechanism [21]. A global descriptor is first obtained by summing the head outputs, $\mathbf{G}_t = \sum_{h=1}^{H} \mathbf{O}_t^h$. This descriptor is passed through a learnable weight matrix $\mathbf{W} \in R^{d_k\times H}$ to generate normalized head-wise importance weights $\alpha_t = \sigma(\mathbf{G}_t\mathbf{W}) \in R^{M\times H}$, where $\sigma(\cdot)$ is the Sigmoid activation function. The final spatial representation at frame t is computed as:

$$\mathbf{Z}_t^{\mathrm{spat}} = \sum_{h=1}^{H} \alpha_t^h \mathbf{O}_t^h, \quad \mathbf{Z}_t^{\mathrm{spat}} \in R^{M\times F}. \tag{5}$$

Then the intra-channel spatial representation $\mathbf{Z}^{\mathrm{spat}}$ is obtained by stacking $\mathbf{Z}_t^{\mathrm{spat}}$ along time frame dimension as $\mathbf{Z}^{\mathrm{spat}} = \{\mathbf{Z}_t^{\mathrm{spat}}\}_{t=1}^{T}$.

TABLE I

EVALUATION OF PESQ, STOI, AND SI-SDRI (DB) ON AD-HOC ARRAYS. DIFFERENT DEST-BASED CONFIGURATIONS ARE COMPARED WITH TAC AND FASTMNMF ACROSS VARYING MICROPHONE COUNTS, OVERLAP RATIOS, AND REVERBERATION TIMES (RT_{60}). * INDICATES A STATISTICALLY SIGNIFICANT DIFFERENCE FOR $p < 0.05$ BETWEEN DCA+IAFF AND TAC, DETERMINED BY T-TEST.

Method	**Params.**	**Overall Metrics**			**Mics (SI-SDRi)**			**Overlap (%) (SI-SDRi)**				**RT_{60} (s) (SI-SDRi)**			
		PESQ	**STOI**	**SI-SDRi**	**2**	**4**	**6**	**0-25**	**25-50**	**50-75**	**75-100**	**0.1-0.2**	**0.2-0.3**	**0.3-0.4**	**0.4-0.5**
FastMNMF	-	2.38	0.825	7.18	6.83	7.38	7.20	7.27	7.29	7.15	6.98	7.15	7.31	6.85	7.42
TAC	2.76M	2.37	0.820	10.98	10.16	11.14	11.57	14.98	11.61	9.56	7.68	11.92	11.48	10.56	9.91
Different DeST-based Configurations															
Mean + Concat	2.62M	2.48	0.835	11.79	11.04	11.93	12.36	15.88	12.42	10.32	8.47	12.93	12.33	11.28	10.58
Mean + iAFF	2.62M	2.49	0.835	11.81	10.95	12.00	12.44	15.88	12.41	10.35	8.51	13.00	12.30	11.30	10.57
MHSA + iAFF	2.62M	**2.52**	**0.841**	12.13	11.18	12.29	12.79	16.04	**12.74**	10.71	8.95	**13.48**	12.63	11.53	10.82
StdCA + iAFF	**2.56M**	2.51	**0.841**	12.17	**11.30**	12.35	12.77	**16.16**	12.72	10.73	9.01	13.44	12.68	**11.61**	10.91
DCA + iAFF	**2.56M**	**2.52***	**0.841***	**12.19***	11.29*	**12.43***	**12.86***	16.08*	**12.74***	**10.78***	**9.06***	**13.48***	**12.70***	11.56*	**10.95***

Because the convolutional projections are applied independently to each channel with shared parameters, and inter-channel interactions are computed solely through symmetric self-attention over the M channels without imposing any channel ordering, the DCA module contains no parameters whose dimensionality depends on M. Consequently, the architecture naturally accommodates an arbitrary number of microphones at inference time, including channel counts unseen during training, and preserves strict permutation invariance [22]. Note that this fixed padding is an implementation choice for batching efficiency rather than an architectural constraint; the network itself contains no dimension tied to the channel count.

B. Temporal Path

The temporal path $\mathcal{T}$ applies a shared S-TCN block [15], [16] to each channel temporal feature $\mathbf{Z}_m$ independently.

Specifically, the S-TCN consists of R stacked squeezed temporal convolution modules (S-TCMs) [16]. Let $\mathbf{Z}_m^{(r)}$ denote the input to the r-th S-TCM of the m-th channel, with $\mathbf{Z}_m^{(1)} = \mathbf{Z}_m$. Each S-TCM refines the representation via a dual-branch gated mechanism:

$$\begin{aligned} \mathbf{Z}_m^{(r)\prime} &= \mathrm{Conv}(\mathbf{Z}_m^{(r)}), \\ \mathbf{Z}_m^{(r)\prime\prime} &= \text{D-Conv}^{(r)}(\mathbf{Z}_m^{(r)\prime}) \otimes \sigma\big(\text{D-Conv}^{(r)}(\mathbf{Z}_m^{(r)\prime})\big), \\ \mathbf{Z}_m^{(r+1)} &= \mathrm{Conv}(\mathbf{Z}_m^{(r)\prime\prime}) + \mathbf{Z}_m^{(r)}, \end{aligned} \tag{6}$$

where $\text{D-Conv}^{(r)}$ denotes a 1-D dilated convolution with a dilation rate 2^{r-1} and $\otimes$ denotes element-wise multiplication.

This dual-branch design, leveraging exponentially increasing dilation rates and a gating mechanism, allows the network to effectively capture long-range temporal dependencies and select salient acoustic features across multiple scales.

After R stacked S-TCMs, the temporal mapping $\mathcal{T}$ is defined as:

$$\mathbf{Z}_m^{\mathrm{temp}} = \mathcal{T}(\mathbf{Z}_m) = \mathbf{Z}_m^{(R+1)}. \tag{7}$$

Finally, the intra-channel temporal features are aggregated as $\mathbf{Z}^{\mathrm{temp}} = \{\mathbf{Z}_m^{\mathrm{temp}}\}_{m=1}^{M}$.

Symmetrically to the DCA, the S-TCN refines the intra-channel temporal representations with serial dilated convolutions; while the DCA module operates as a learnable multi-resolution filter bank, and leverages parallel dilated convolutions to compute the inter-channel spatial representations, capturing global correlations across microphones at varying temporal resolutions.

C. Fusion

The fusion module iAFF, denoted as $\mathcal{F}$, performs a two-step gated fusion using the multi-scale channel attention module (MS-CAM) [17], denoted as $\mathcal{M}(\cdot)$.

Step 1 (Coarse Fusion): An initial attention weight is computed by

$$w = \mathcal{M}(\mathbf{Z}^{\mathrm{spat}} + \mathbf{Z}^{\mathrm{temp}}), \tag{8}$$

which is used to obtain an intermediate representation:

$$\mathbf{Z}' = w \otimes \mathbf{Z}^{\mathrm{spat}} + (1 - w) \otimes \mathbf{Z}^{\mathrm{temp}}. \tag{9}$$

Step 2 (Refinement): The attention is recalculated based on the intermediate feature,

$$\hat{\mathbf{Z}} = \mathcal{M}(\mathbf{Z}') \otimes \mathbf{Z}^{\mathrm{spat}} + \big(1 - \mathcal{M}(\mathbf{Z}')\big) \otimes \mathbf{Z}^{\mathrm{temp}}. \tag{10}$$

This iterative design reweights intra-channel temporal features based on inter-channel spatial features, enhancing channel-aware modeling.

III. EXPERIMENTAL SETUP

A. Dataset Generation

We consider the two-speaker separation scenario ($I = 2$). For each speaker, clean speech utterances are randomly sampled from LibriSpeech [23], following the data configuration in [12].

Room impulse responses (RIRs) are generated using gpuRIR [24]. Room dimensions are uniformly sampled from $[3, 10] \times [3, 10] \times [2.5, 4]$ m with $\mathrm{RT}_{60} \in [0.1, 0.5]$ s, and both speakers are randomly positioned within the room. Their speech signals are spatialized by convolution with the corresponding RIRs and then mixed with an overlap ratio uniformly distributed in 0–100% and a relative SNR between the two speakers uniformly sampled from $[0, 5]$ dB. Diffuse-like noise is simulated by placing 1–30 random point sources, which are

spatialized using the same RIR generation process and added to the mixture at an SNR uniformly sampled from $[-5, 5]$ dB. All SNRs are defined on the dry source signals at their emission positions, i.e., the gains are applied to the source signals prior to RIR convolution. Consequently, the local SNR observed at each microphone is not directly controlled but emerges from the acoustic propagation (source–microphone distances and reverberation), which reflects realistic ad-hoc deployment conditions where microphone utility varies across channels.

Experiments are conducted on both ad-hoc and fixed microphone arrays: *1)* ***Ad-hoc array:*** 2–6 microphones randomly placed within the room; *2)* ***Fixed array:*** a 4-channel circular array (radius 3.2 cm). For each array configuration, we generate $20\text{k}/5\text{k}/3\text{k}$ training/validation/test samples (4 s, 16 kHz).

B. Implementation Details

The FaSNet backbone stacks four blocks ($P = 4$), each followed by a DeST block, resulting in four DeST modules. Within each DeST, the feature dimension is expanded to $F = 128$. The temporal path (S-TCN) uses $R = 6$ modules, and the spatial path adopts $H = 4$ heads with dilation rates $1/2/4/8$ and projection dimension $d_k = 128$.

To form mini-batches with varying channel counts, mixtures with fewer microphones are zero-padded along the channel dimension to a fixed count of 6 channels during both training and inference. A binary channel mask is applied so that padded channels are excluded from the softmax in Eq. (4) and from the fusion module, and their outputs are discarded before decoding; hence padded channels do not contribute to the separated signals or the training loss.

All models are trained by minimizing the negative SI-SDR [25] with Permutation Invariant Training (PIT) [26] and the AdamW optimizer [27]. The learning rate is set to 5×10^{-4} with a weight decay of 1×10^{-6}. Performance is evaluated in terms of SI-SDR improvement (SI-SDRi), Perceptual Evaluation of Speech Quality (PESQ) [28], and Short-Time Objective Intelligibility (STOI) [29].

IV. Results and Analysis

A. Performance on Ad-hoc Arrays

Table I reports the evaluation on ad-hoc arrays. We compare the proposed method with FastMNMF [3], an iterative EM-based method that naturally supports an arbitrary number of microphones and is invariant to channel permutations, and TAC.

For FastMNMF [3], we follow the standard configuration: the diagonalizer is initialized with the identity matrix, the NMF source parameters are randomly initialized, and the number of NMF bases is set to K = 16 per source. The EM-style updates are run for 100 iterations per mixture.

To investigate the effectiveness of the spatial–temporal decoupling, we evaluate several configurations of the DeST architecture. For the spatial path, four alternatives are considered: mean pooling (**Mean**), **MHSA**, standard convolution attention without dilation (**StdCA**), and the proposed **DCA** described in Section II-A. For the temporal path, we adopt the S-TCN introduced in Section II-B, shared by all variants and omitted in Table I. Two fusion strategies are evaluated: concatenation (**Concat**) and the **iAFF** described in Section II-C. The final **proposed** configuration corresponds to **DCA+iAFF**.

TABLE II
SI-SDRi (dB) evaluation on Fixed Geometry Arrays with FaSNet backbone.

Backbone	Method	Params.	SI-SDRi
FaSNet	TAC	2.76M	10.76
	Mean + Concat	2.62M	11.35
	MHSA + iAFF	2.62M	11.37
	DCA + iAFF	**2.56M**	**12.02**

TABLE III
SI-SDRi (dB) evaluation on Fixed Geometry Arrays with IC-ConvTasNet backbone.

Backbone	Spatial Path	Params.	SI-SDRi
IC-ConvTasNet	–	0.80M	8.34
	Mean	0.80M	8.54
	MHSA	1.00M	8.80
	DCA	**0.83M**	**8.96**

The results show that the DeST-based methods achieve improved performance over both FastMNMF and TAC across different microphone counts, overlap ratios, and RT_{60}, with the proposed configuration yielding the best results, demonstrating the effectiveness of spatial–temporal decoupling for ad-hoc speech separation. Performance generally improves with an increasing number of microphones, as richer spatial information becomes available. In contrast, higher overlap ratios and larger RT_{60} values increase the difficulty of the separation task, leading to reduced performance.

B. Generalization to Fixed Geometry Arrays

To evaluate generalization to fixed-geometry arrays, we compare TAC and DeST-based methods (Mean+Concat, MHSA+iAFF and the proposed DCA+iAFF) on the fixed-array test set (Table II). The proposed DCA+iAFF achieves 12.02 dB, substantially surpassing the TAC (10.76 dB). Specifically, the MHSA-based DeST suffers a larger performance drop (from 12.13 dB to 11.37 dB) on the fixed array, falling to a level similar to the Mean+Concat variant (11.35 dB). In contrast, the DCA-based variant remains stable, with only a marginal decrease from 12.19 dB to 12.02 dB. We attribute this to the compact 3.2 cm array radius, which results in highly correlated microphone signals and limited spatial diversity. Standard MHSA employs single-scale linear projections that struggle to resolve such subtle spatial differences, causing attention weights to degenerate toward uniform averaging. In contrast, the multi-resolution DCA expands the temporal context and accumulates subtle phase variations over time, enabling more robust attention estimation.

To further verify the generalizability of the DeST across different separation backbones, we additionally adopt IC-ConvTasNet as the backbone [30]. Since IC-ConvTasNet already adopts TCN for temporal modeling, we omit the

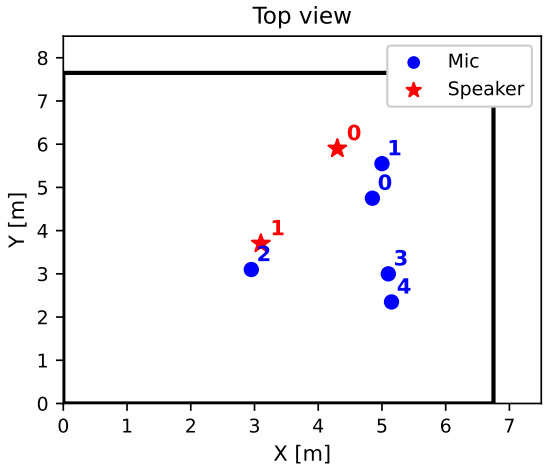
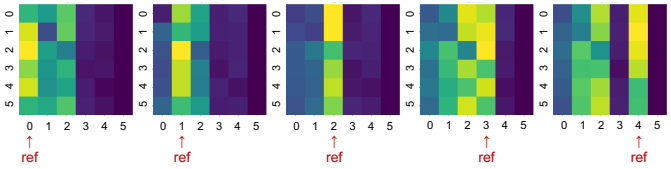
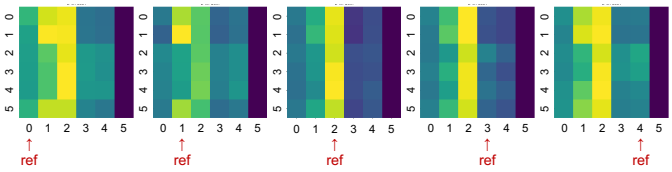


Fig. 3. Spatial configuration of the case study. Speaker 0 is closest to Microphone 1, and Speaker 1 to Microphone 2.

temporal path and only evaluate the proposed spatial path, including three variants: Mean pooling, MHSA, and DCA. As shown in Table III, introducing explicit spatial modeling consistently improves the IC-ConvTasNet baseline (8.34 dB), with DCA achieving the best performance (8.96 dB), further demonstrating the effectiveness of decoupled spatial–temporal modeling.



(a) MHSA + iAFF



(b) DCA + iAFF

Fig. 4. Visualization of Layer-2 attention weights under different reference microphone assignments (Channels 0–4, red arrows). The x-axis denotes the Key (attended channel) and the y-axis the Query; lighter colors indicate higher weights. Channel 5 corresponds to the zero-padded channel; its attention weights are masked out and shown for completeness.

C. Visualization Analysis

We analyze a 5-channel ad-hoc case study, whose top-view spatial configuration is shown in Fig. 3. Correspondingly, Fig. 4 visualizes the Layer-2 attention weights under different reference microphone assignments for MHSA + iAFF and the proposed DCA + iAFF. Each row shows the attention of one microphone to the others, where the visualized weights are those used in Eq. (4) to aggregate information from other channels into each channel. We interpret the learned attention weights as an implicit estimate of microphone utility. As shown in Fig. 4a, MHSA + iAFF exhibits severe "reference bias": a prominent vertical stripe shifts with the assigned reference index, regardless of the actual SNR of the sensors. In contrast, Fig. 4b shows that the multi-scale DCA maintains spatial stability: even when a low-SNR microphone is selected as the reference, it adaptively concentrates attention on high-SNR microphones closer to the speakers, demonstrating effective microphone utility modeling.

V. Conclusion

In this paper, we proposed a DeST block for ad-hoc multi-channel speech separation. The proposed design explicitly separates spatial and temporal modeling through parallel paths, where DCA captures channel-aware spatial features and an S-TCN extracts intra-channel temporal features. These features are integrated through iAFF-based fusion to achieve utility-aware feature aggregation. Experimental results on simulated ad-hoc arrays demonstrate consistent improvements over TAC and FastMNMF across various microphone counts, overlap ratios, and reverberation conditions. Additional evaluations on fixed-geometry arrays and different separation backbones further confirm the robustness and generalizability of the proposed design. Future work will explore extending the learned spatial representations for speaker localization in ad-hoc microphone arrays, as the learned attention patterns tend to emphasize microphones closer to the speakers with higher SNR.

References

[1] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: Two converging routes to ILRMA originating from ICA and NMF," *APSIPA Transactions on Signal and Information Processing*, vol. 8, e12, 2019.

[2] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 196–200.

[3] K. Sekiguchi, A. A. Nugraha, Y. Bando, and K. Yoshii, "Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices," in *Proceedings of the 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[4] Y. Wang et al., "A study of multichannel spatiotemporal features and knowledge distillation on robust target speaker extraction," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 431–435.

[5] Y. Wang, J. Zhang, C. Jiang, W. Zhang, Z. Ye, and L. Dai, "Leveraging boolean directivity embedding for binaural target speaker extraction," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[6] Y. Wang, J. Chen, W. Zhou, and T. Kawahara, "Position-aware target speaker extraction for long-form multi-party conversations: A diarization-free framework for asr," *arXiv preprint arXiv:2606.29497*, 2026.

[7] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[8] K. Kinoshita, M. Souden, and T. Nakatani, "Blind source separation using spatially distributed microphones based on microphone-location dependent source activities," in *Proc. INTERSPEECH*, 2013, pp. 822–826.

[9] X.-L. Zhang, "Deep ad-hoc beamforming," *Computer Speech & Language*, vol. 68, p. 101 201, 2021.

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

[11] H. Taherian, S. E. Eskimez, T. Yoshioka, H. Wang, Z. Chen, and X. Huang, "One model to enhance them all: Array geometry agnostic multi-channel personalized speech enhancement," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 271–275.

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

[13] J. Fan, R. Gu, Y. Luo, and C. Pang, "A unified geometry-aware source localization and separation framework for AD-HOC microphone array," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 725–729.

[14] T. C. Lawin-Ore and S. Doclo, "Reference microphone selection for MWF-based noise reduction using distributed microphone arrays," in *Proceedings of the 10th ITG Symposium on Speech Communication*, 2012, pp. 1–4.

[15] A. Li, W. Liu, C. Zheng, and X. Li, "Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6487–6491.

[16] A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1829–1843, 2021.

[17] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 3560–3569.

[18] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 260–267.

[19] D. Lee and J.-W. Choi, "DeFTAN-AA: Array geometry agnostic multichannel speech enhancement," in *Proc. INTERSPEECH*, 2024, pp. 3360–3364.

[20] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[21] W. Ren et al., "Gated fusion network for single image dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Recognition (CVPR)*, 2018, pp. 3253–3261.

[22] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep sets," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[24] 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, no. 4, pp. 5653–5671, 2021.

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

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

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[28] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 2001, pp. 749–752.

[29] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on audio, speech, and language processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[30] D. Lee, S. Kim, and J.-W. Choi, "Inter-channel Conv-TasNet for multichannel speech enhancement," *arXiv preprint arXiv:2111.04312*, 2021.