

SUBBAND-WISE ATTENTION-BASED NEURAL MVDR BEAMFORMING FOR DYNAMIC SOURCES

Kyeongmuk Kang and Jong Won Shin

Gwangju Institute of Science and Technology, Gwangju, 61005, South Korea
(e-mail: kyeongmuk@gm.gist.ac.kr; jwshin@gist.ac.kr)

ABSTRACT

Fast and accurate estimation of the spatial covariance matrices (SCMs) for speech and noise is crucial for spatial filtering to enhance noisy speech when sound sources are moving. Self-attention-based time-varying minimum variance distortionless response (MVDR) filter (ATT-MVDR) was proposed which aggregates frequency-wise instantaneous SCMs (ISCMs) across frames using attention weights for those frames. While the idea of summarizing ISCMs for only relevant frames was effective, the same weight is applied to all frequency bands for each frame, although the local signal-to-noise ratios can differ across frequencies. In this paper, we propose a subband-wise attention-based neural MVDR filter, which determines the attention weights for each subband separately to deal with the cases when the spatial information is reliable in only a subset of subbands. We have also adopted a multi-channel mask estimator based on the Embedding-based Mask Network (EMNet) to exploit spatial information also in the mask estimation stage. Experiments on moving-source mixtures showed the proposed method achieved comparable or better speech quality and automatic speech recognition performance with a smaller number of parameters.

Index Terms— Speech enhancement, MVDR filter, spatial covariance matrix, moving speech sources, self-attention

1. INTRODUCTION

One of the representative techniques to enhance noisy speech signals from multiple microphones is beamforming, which applies a linear spatial filter optimizing a certain criterion [1, 2]. Although linear spatial filtering such as the minimum variance distortionless response (MVDR) beamforming cannot suppress the interference signals more than a certain level, it does not introduce any nonlinear distortion to the desired

speech and thus is useful for some applications such as automatic speech recognition (ASR) [3, 4, 5]. The spatial filter is often determined by spatial covariance matrices (SCMs) for desired speech and noise. Recent approaches utilized deep neural networks (DNNs) to estimate parameters for the spatial filters or post-filters following them. Unlike the end-to-end neural multi-channel speech enhancement (MCSE) which directly estimates clean speech from multi-channel noisy observations [6, 7, 8], DNN-based MCSE approaches with spatial filters adopted the DNNs to estimate masks to obtain signals necessary for the SCMs estimation [5, 9, 10, 11, 12], acoustic or statistical parameters for the parameter estimation [13, 14, 15], or the beamformer coefficients directly [16, 17]. The estimation of the SCMs inevitably involves a certain form of the temporal aggregation of the statistics over multiple frames, which may not be straightforward for moving source scenarios. Conventional recursive or block-wise approaches estimate time-varying SCMs using a predefined forgetting factor or block size. In [18], the forgetting factor is estimated along with the time-frequency masks, but still the statistics in the distant frames with the same spatial location cannot be utilized.

To address this limitation, an attention-based MVDR filter (ATT-MVDR) [11] was proposed, which aggregates spatial cues in the time frames with self-attention weights and thus can deal with moving sources. However, the spectral masks to estimate instantaneous SCM estimates (ISCMs) are based on each individual microphone signals, and more importantly, the same attention weight is applied to all frequency bins in a certain frame although the signal-to-noise ratios (SNRs) in the frequency bins can differ significantly. Because speech and many of the interfering signals are sparse in the time-frequency (TF) domain, the spatial information for a source may be preserved in some frequency bins while it is deteriorated in other bins for the same time frame.

A few attempts have been made to improve the ATT-MVDR. In [12], nonlinear attention-based MVDR filters in which the SCMs (NL-ATT-MVDR) or the scaled inverse of the SCMs (IC-MVDR) are directly estimated by the neural networks were compared with the ATT-MVDR. ABIC-MVDR was proposed in [4] which adopts a single network to estimate both the masks to compute ISCMs and

This work was partly supported by the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2026-RS-2021-II211835, 0%), and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No.IRIS-2025-25431727, 0%).

the frequency-wise attention weights from the complex short-time Fourier transform (STFT) coefficients for multiple microphone signals. While it alleviates main drawbacks of the ATT-MVDR, the attention weights are computed directly from the multi-mic signals instead of ISCMs which may make it difficult to put higher weights for the frames with similar spatial characteristics.

In this paper, we propose a subband-wise attention-based neural MVDR filter (SB-ATT-MVDR) in which the subband-wise attention weights are estimated from the ISCMs obtained using multichannel Embedding-based Mask Network (EM-Net). Subband-wise attention enables band-selective attention to the TF bins in which spatio-spectral cues for the source are retained and also reduces the number of parameters for the SCM estimation, while the multichannel mask estimator can exploit spatial diversity of source positions. Experimental results showed that the proposed SB-ATT-MVDR demonstrated comparable or better speech quality and ASR performance with only 2.7M parameters.

2. THE ATTENTION-BASED MVDR FILTER

We consider a dynamic acoustic environment in which the target source moves within an utterance. The noisy signals captured by C microphones in the TF domain, $\mathbf{Y}(t, f) \in \mathbb{C}^{C \times 1}$, are modeled as

$$\mathbf{Y}(t, f) = \mathbf{H}(t, f)S(t, f) + \mathbf{N}(t, f), \quad (1)$$

where $\mathbf{H}(t, f) \in \mathbb{C}^{C \times 1}$ denotes time-varying acoustic transfer functions (ATFs) from the position of the desired speaker at time t to the microphones, and $S(t, f)$ and $\mathbf{N}(t, f)$ represent the complex STFT coefficients of the clean speech and noise including directional interferences, respectively. The goal of the linear spatial filtering is to find a filter $\mathbf{w}(t, f) \in \mathbb{C}^{C \times 1}$ which is applied to $\mathbf{Y}(t, f)$ to estimate $S(t, f)$, i. e.,

$$\hat{S}(t, f) = \mathbf{w}^H(t, f)\mathbf{Y}(t, f). \quad (2)$$

Souden et al. [19] proposed the MVDR filter given by

$$\mathbf{w}_{\text{MVDR}}(t, f) = \frac{(\Phi^{\mathbf{N}}(t, f))^{-1}\Phi^{\mathbf{S}}(t, f)}{\text{Tr}[(\Phi^{\mathbf{N}}(t, f))^{-1}\Phi^{\mathbf{S}}(t, f)]}\mathbf{u}, \quad (3)$$

where $\Phi^{\mathbf{S}}(t, f) \in \mathbb{C}^{C \times C}$ and $\Phi^{\mathbf{N}}(t, f) \in \mathbb{C}^{C \times C}$ represent the SCMs of the desired signal and noise, respectively, and $\mathbf{u} \in \mathbb{C}^C$ is a one-hot vector corresponding to the reference microphone, which is equivalent to the MVDR beamformer when $\Phi^{\mathbf{S}}(t, f)$ is of rank one. Mask-based MVDR filters [9, 10, 20, 11, 12] utilize DNNs to estimate TF masks required to obtain signal and noise SCMs. The ISCM $\Psi^\nu(t, f)$ for the source ν , which is either desired speech or noise, can be expressed as

$$\Psi^\nu(t, f) = m^\nu(t, f)\mathbf{Y}(t, f)\mathbf{Y}(t, f)^H, \quad (4)$$

where $m^\nu(t, f)$ is the mask for the source ν estimated by a DNN. Then, the SCM for the source ν , $\Phi^\nu(t, f)$, can be obtained as a weighted summation of the ISCMs:

$$\Phi^\nu(t, f) = \sum_{t'=1}^T A_f^\nu(t, t')\Psi^\nu(t', f), \quad (5)$$

in which $A_f^\nu(t, t')$ denotes the contribution of the ISCM for the frame t' to the SCM estimate at the frame t and the frequency bin f .

Unlike conventional schemes that predetermine $A_f^\nu(t, t') = A^\nu(t, t')$ so that (5) becomes recursive smoothing or moving average, ATT-MVDR [11] computes frame-wise weights applied to all frequencies as a function of the input data using the attention mechanism. The ISCMs $\Psi^\nu(t, f) \in \mathbb{C}^{C \times C}$ in all F frequencies are first flattened to form a $2FC^2$ dimensional real-valued vector $\psi^\nu(t)$ which is then squeezed into a D -dimensional vector $E^{\nu,0}(t)$ by a single linear layer. $E^{\nu,0} = \{E^{\nu,0}(t)\}_{t=1}^T \in \mathbb{R}^{T \times D}$ is then processed by I Transformer encoder layers.

$$\psi^\nu(t) = \text{Flatten}(\{\Psi^\nu(t, f)\}_{f=1}^F) \in \mathbb{R}^{2FC^2}, \quad (6)$$

$$E^{\nu,0}(t) = \text{Linear}(\psi^\nu(t)) \in \mathbb{R}^D. \quad (7)$$

The encoder outputs are then passed to an attention module; we denote the whole mapping by $\text{NN}(\cdot; \Lambda^\nu)$ with trainable parameters Λ^ν , yielding the temporal attention matrix $\mathbf{A}^\nu = \text{NN}(\{\psi^\nu(t)\}_{t=1}^T; \Lambda^\nu) \in \mathbb{R}^{T \times T}$; its weights are applied in (5) to obtain the speech and noise SCMs, from which the ATT-MVDR filter follows via (3).

3. PROPOSED SUBBAND-WISE ATTENTION-BASED MVDR BEAMFORMER

As discussed in Section I, ATT-MVDR has two key limitations: single-channel mask estimation and frequency-agnostic temporal attention weights. The latter is particularly limiting because frequency-domain MVDR is a narrowband beamformer, whereas ATT-MVDR applies a single temporal attention matrix to all frequency bins. Since local target-to-noise dominance, and thus the temporal frames useful for SCM aggregation, can vary across frequency regions, SCM estimation benefits from frequency-dependent temporal weighting. We therefore generalize attention-based SCM estimation by estimating temporal attention weights separately for subbands while applying them to frequency-wise ISCMs, and instantiate this framework as the subband-wise attention-based neural MVDR filter (SB-ATT-MVDR). Furthermore, we propose the Embedding-based Mask Network (EMNet), which extends the single-channel mask estimation to multi-channel through an embedding extraction module that captures inter-channel spatial information. The effectiveness of the proposed SB-ATT-MVDR is validated in Section IV.

Table 1: Objective evaluation results on the moving source dataset.

Model	Param. (M)	GMAC/s	PESQ-nb	PESQ-wb	STOI	SDR	WER
Noisy	-	-	1.90	1.34	0.83	4.93	28.07
ConvTasNet	8.8	2.84	2.93	2.31	0.93	13.38	16.93
- ATT-MVDR	38.0	4.73	2.98	2.58	0.95	15.98	11.72
- NL-ATT-MVDR	40.6	4.87	3.15	2.63	0.95	15.45	13.96
- IC-MVDR	40.6	4.87	3.13	2.65	0.95	15.41	13.87
- SB-ATT-MVDR	10.2	5.01	3.17	2.73	0.96	16.17	12.03
ABIC-MVDR	0.35	4.01	3.20	2.78	0.97	16.68	8.21
EaBNet	2.83	7.35	3.28	2.86	0.97	16.31	7.94
EMNet	1.28	3.04	3.35	2.84	0.96	15.58	7.76
- ATT-MVDR	30.5	4.98	3.16	2.82	0.97	17.29	7.49
- NL-ATT-MVDR	33.1	5.11	3.42	2.96	0.97	17.22	7.97
- IC-MVDR	33.1	5.11	3.37	2.94	0.97	16.97	7.96
- SB-ATT-MVDR	2.7	5.25	3.34	2.97	0.97	17.43	7.69

3.1. The generalized attention-based SCM estimation

Let N denote the number of subbands and G_n the number of frequency bins in subband n . With $B_0 = 0$ and $B_n = \sum_{k=1}^n G_k$, $\mathcal{F}_n = \{B_{n-1} + 1, \dots, B_n\}$ denotes the frequency bins assigned to subband n , $n = 1, \dots, N$. Building upon (5), we constrain the frequency-indexed weight to be shared within each subband, $A_f^\nu(t, t') = A_n^\nu(t, t')$ for $f \in \mathcal{F}_n$, and estimate one attention matrix per subband. The spatial embedding of subband n at frame t is defined as:

$$\psi_n^\nu(t) = \text{Flatten}(\{\Psi^\nu(t, f)\}_{f \in \mathcal{F}_n}) \in \mathbb{R}^{2G_n C^2}, \quad (8)$$

$$E_n^{\nu,0}(t) = \text{Linear}_n(\psi_n^\nu(t)) \in \mathbb{R}^D \quad (9)$$

This subband projection reduces the input dimension of each linear layer from the full-band size $2FC^2$ in (7) to $2G_n C^2$. For each subband n , the same estimator produces $\mathbf{A}_n^\nu \in \mathbb{R}^{T \times T}$:

$$\mathbf{A}_n^\nu = \text{NN}(\{\psi_n^\nu(t)\}_{t=1}^T; \mathbf{A}^\nu) \in \mathbb{R}^{T \times T}. \quad (10)$$

Substituting the subband-sharing constraint into (5), the final SCM is obtained as:

$$\Phi^\nu(t, f) = \sum_{t'=1}^T A_n^\nu(t, t') \Psi^\nu(t', f), \quad f \in \mathcal{F}_n. \quad (11)$$

Setting $N = 1$ with $\mathcal{F}_1 = \{1, \dots, F\}$ reduces this formulation to the SCM estimator of ATT-MVDR.

3.2. Subband-wise Attention-based MVDR beamformer

SB-ATT-MVDR uses subband-specific temporal attention weights to aggregate frequency-wise ISCMs before deriving the MVDR beamforming weights. For 16 kHz sampled signals, the frequency bins are grouped into three ranges: low (0–2 kHz, $G_n = 8$), mid (2–4 kHz, $G_n = 16$), and high (4–8 kHz, $G_n = 32$), resulting in $N = 16 + 8 + 8 = 32$ subbands. This subband configuration reflects that local target-to-noise

dominance can vary across frequency regions. We therefore use smaller G_n values in lower frequency regions and larger G_n values in higher frequency regions, yielding finer subbands at low frequencies and coarser subbands at high frequencies. Within each frequency range, subbands with the same G_n share the corresponding linear projection layer. The effectiveness of this configuration is validated in Section IV-C.

3.3. Embedding-based mask estimator

A fundamental element of mask-based beamforming involves the accurate estimation of speech and noise masks from multi-channel observations, which facilitates the derivation of reliable spatial statistics. Specifically, an all-neural causal DNN-based beamformer introduced in [16] for multi-channel speech enhancement has demonstrated robust performance across diverse SNR conditions. This architecture incorporates an Embedding Module (EM), which utilizes an "Encoder-TCN-Decoder" topology to encode multi-resolution features into a 3D embedding tensor, along with a Beamforming Module (BM) to compute framewise weights. In the proposed framework, we leverage the EM, which is designed to encode multi-channel spatial information, to transform multi-channel noisy speech into low-dimensional 3D embeddings that capture essential target speech characteristics. Subsequently, the target speech mask is estimated via the mask decoder architecture adapted from CMGAN [21], which has demonstrated strong performance in single-channel SE. By processing multi-channel inputs, EMNet captures inter-channel spatial cues such as directional information of the source, producing spatially informed masks that better reflect source movement compared with single-channel mask estimators. Formally, EMNet takes the multi-channel observation $\mathbf{Y} \in \mathbb{C}^{C \times T \times F}$ as input and estimates the reference-channel speech mask $\hat{m}^\nu(t, f) = \text{EMNet}(\mathbf{Y}; \Theta)$, where Θ denotes the trainable parameters.

Table 2: Objective results for the moving datasets with causal settings.

Model	PESQ-nb	PESQ-wb	STOI	SDR	WER
Noisy	1.90	1.34	0.83	4.93	28.07
EaBNet	3.13	2.71	0.96	16.06	8.30
ABIC-MVDR	3.17	2.73	0.97	16.02	8.71
EMNet	3.27	2.75	0.96	15.32	8.53
- ATT-MVDR [11]	3.05	2.62	0.96	15.98	7.97
- NL-ATT-MVDR [12]	3.35	2.89	0.97	16.92	8.29
- IC-MVDR	3.29	2.85	0.97	16.63	8.32
- SB-ATT-MVDR	3.27	2.85	0.97	16.85	7.69

4. EXPERIMENTS

4.1. Baselines and model configuration

Following [11], the baseline mask estimator is a modified ConvTasNet [22] using STFT/iSTFT [23] for mask estimation. We employ Instance Normalization (IN) for the experiments reported here. SB-ATT-MVDR uses $I = 2$ Transformer encoder layers with four attention heads; full architectures, STFT settings, causal configurations, and corresponding results are provided on the demo page. ATT-MVDR [11], NL-ATT-MVDR, and IC-MVDR [12] serve as attention-based baselines. ABIC-MVDR [4] and EaBNet [16] are included as lightweight attention-based and end-to-end neural beamforming baselines, respectively.

4.2. Experimental configuration

Building upon [11, 12, 4], we simulated moving-source mixtures using gpuRIR [24] with WSJ0 speech [25] and a 5-channel CHIME-3 microphone configuration [26]. Unlike prior attention-based MVDR studies, we introduced a moving competing talker in addition to diffuse background noise, with the target-to-interference-plus-noise ratio (SINR) ranging from 0 to 10 dB and the interferer-to-diffuse-noise ratio (INR) ranging from -10 to 10 dB. Following [11], all models except EaBNet minimized the scale-dependent SNR loss [27] between the target and enhanced waveforms; EaBNet instead used its original combined spectral-magnitude and real-imaginary MSE loss [16]. All models were trained with Adam and early stopping; detailed room geometry, source trajectories, SINR and INR settings, and training configurations are provided on the demo page.

4.3. Experimental results

To evaluate the proposed approach, we trained all models on simulated moving source datasets, as summarized in Table 1. We used Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI), Signal-to-Distortion Ratio (SDR), and Word Error Rate (WER) cal-

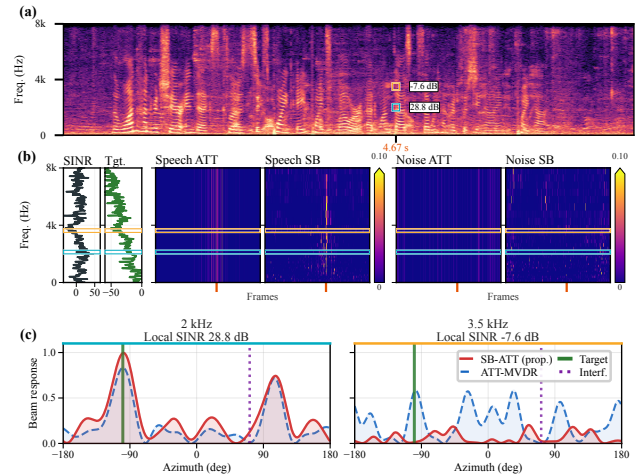


Fig. 1: Attention weights and beampatterns of ATT-MVDR and SB-ATT-MVDR at $t = 4.67$ s: (a) noisy spectrogram with the analyzed time-frequency points, (b) local SINR, target STFT magnitude, and speech/noise attention weight maps, and (c) beampatterns at 2 and 3.5 kHz.

culated using Whisper-tiny [28] for evaluation, and compared trainable parameters and GMACs for efficiency. Experimental results demonstrate that EMNet, as a standalone mask estimator without beamforming, consistently outperforms the ConvTasNet-based mask estimator across all objective measures with fewer parameters. While ConvTasNet struggles to suppress interference from competing speech—leading to a persistently high WER—EMNet suppresses the interfering talker and background noise while preserving the moving target speech under dynamic acoustic conditions.

To evaluate the effect of spatial filtering, the mask estimator was alternated between ConvTasNet and EMNet under identical beamforming backends (Table 1). The proposed SB-ATT-MVDR integrated with EMNet achieves the highest PESQ-wb and SDR with only 2.7M parameters, compared with 30.5M for the EMNet-integrated ATT-MVDR. Although its WER is slightly higher than that of EMNet-integrated ATT-MVDR (7.69% vs. 7.49%), SB-ATT-MVDR preserves robust enhancement with substantially higher parameter efficiency.

We further evaluated the same attention-based beamforming methods in causal settings, as shown in Table 2. For the causal setting, the normalization in EMNet was changed to cumulative layer normalization, and causal convolution was applied. For attention-based MVDR beamforming methods, a causal mask was applied to the weight coefficients to prevent utilizing future information. EMNet alone achieves competitive PESQ—surpassing the EMNet+ATT-MVDR combination in PESQ—demonstrating that the multi-channel mask estimator effectively captures spatial cues without explicit beamforming. In contrast, SB-ATT-MVDR improves all

metrics over ATT-MVDR while retaining the much smaller parameter count reported in Table 1. Relative to NL-ATT-MVDR, SB-ATT-MVDR gives slightly lower PESQ and SDR, but achieves a lower WER.

Fig. 1 compares ATT-MVDR and SB-ATT-MVDR at two representative time-frequency bins of the frame at $t = 4.67$ s. Unlike ATT-MVDR, which applies a single temporal attention vector across all frequencies, SB-ATT-MVDR estimates subband-specific weights. For example, in the speech-SCM weight maps shown in panel (b), ATT-MVDR exhibits the same temporal-weight pattern at 2 and 3.5 kHz, whereas SB-ATT-MVDR exhibits distinct patterns for the corresponding subbands. For SB-ATT-MVDR, the high local SINR and large target-source STFT magnitude at 2 kHz coincide with a pronounced speech-SCM attention peak at the selected frame; at 3.5 kHz, where both quantities are lower, the selected-frame weight is much smaller, and higher attention weights are assigned to other frames. Correspondingly, at 2 kHz (local SINR: 28.8 dB), the two methods exhibit similar beampatterns, with both showing a strong response toward the target direction, while the proposed method shows a slightly higher gain toward the interference direction. By contrast, at 3.5 kHz (local SINR: -7.6 dB), ATT-MVDR retains a higher response toward the target direction than SB-ATT-MVDR. This contrast illustrates how subband-wise attention allows the beamformer to adapt its spatial response to frequency-dependent target dominance. The default subband configuration, $G_n = [8, 16, 32]$ over 0–2, 2–4, and 4–8 kHz ranges, was selected based on the demo page ablation studies. The page also reports oracle-mask results across the test set: target and noise gains by local SINR range, and upper-bound performance. Temporal attention weights for a source trajectory that revisits the same direction, where distant frames share spatial characteristics, are also shown there.

5. CONCLUSION

In this paper, we proposed SB-ATT-MVDR, which estimates subband-specific temporal attention weights for SCM aggregation. We also introduced EMNet, a multi-channel mask estimator that encodes multi-channel observations into compact spatial embeddings. By avoiding a single full-band temporal attention matrix, SB-ATT-MVDR applies subband-specific temporal weights to frequency-wise ISCM aggregation while reducing the parameter count of the SCM estimators. Experimental results on the moving source datasets show that SB-ATT-MVDR with EMNet achieves strong PESQ, STOI, and SDR scores while maintaining competitive word error rates, using 2.7M trainable parameters with comparable computational complexity.

6. REFERENCES

- [1] Harry L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*, John Wiley & Sons, 2002.
- [2] Jacob Benesty, Jingdong Chen, and Yiteng Huang, *Microphone Array Signal Processing*, Springer, Berlin, Heidelberg, 2008.
- [3] Ui-Hyeop Shin and Hyung-Min Park, “Statistical beamformer exploiting non-stationarity and sparsity with spatially constrained ICA for robust speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4091–4104, 2024.
- [4] Jinglin Bai, Hao Li, Xueliang Zhang, and Fei Chen, “Attention-based beamformer for multi-channel speech enhancement,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
- [5] Runduo Han, Yanxin Hu, Yihui Fu, Zihan Zhang, Yukai Jv, Li Chen, and Lei Xie, “CabinSep: IR-augmented mask-based MVDR for real-time in-car speech separation with distributed heterogeneous arrays,” in *Proc. Interspeech*, 2025, pp. 4978–4982.
- [6] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe, “TF-GridNet: Integrating full- and sub-band modeling for speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.
- [7] Changsheng Quan and Xiaofei Li, “SpatialNet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1310–1323, 2024.
- [8] Changsheng Quan and Xiaofei Li, “Multichannel long-term streaming neural speech enhancement for static and moving speakers,” *IEEE Signal Processing Letters*, vol. 31, pp. 2295–2299, 2024.
- [9] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach, “Neural network based spectral mask estimation for acoustic beamforming,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.
- [10] Takuya Higuchi, Nobutaka Ito, Shoko Araki, Takuya Yoshioka, Marc Delcroix, and Tomohiro Nakatani, “Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

- [11] Tsubasa Ochiai, Marc Delcroix, Keisuke Kinoshita, and Tomohiro Nakatani, "Mask-based neural beamforming for moving speakers with self-attention-based tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 835–848, 2023.
- [12] Yichen Wang, Archontis Politis, and Tuomas Virtanen, "Attention-driven multichannel speech enhancement in moving sound source scenarios," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11221–11225.
- [13] Minseung Kim, Sein Cheong, and Jong Won Shin, "DNN-based parameter estimation for MVDR beamforming and post-filtering," in *Proc. Interspeech*, 2023, pp. 3879–3883.
- [14] Sein Cheong, Minseung Kim, and Jong Won Shin, "Integrated DNN-based parameter estimation for multichannel speech enhancement," *IEEE Signal Processing Letters*, vol. 32, pp. 3320–3324, 2025.
- [15] Shuai Tao, Pejman Mowlae, Jesper Rindom Jensen, and Mads Græsbøll Christensen, "Multi-channel speech enhancement guided by learning-based A Posteriori speech presence probability estimation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 3838–3849, 2025.
- [16] Andong Li, Chengshi Zheng, Lianwu Zhang, and Xiaodong Li, "Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.
- [17] Hansol Kim, Kyeongmuk Kang, and Jong Won Shin, "Factorized MVDR deep beamforming for multichannel speech enhancement," *IEEE Signal Processing Letters*, vol. 29, pp. 1898–1902, 2022.
- [18] Masahito Togami, "Simultaneous optimization of forgetting factor and time-frequency mask for block online multi-channel speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [19] Mehrez Souden, Jacob Benesty, and Sofiene Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, 2009.
- [20] Xiong Xiao, Shengkui Zhao, Douglas L. Jones, Eng Siong Chng, and Haizhou Li, "On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.
- [21] Sherif Abdulatif, Ruizhe Cao, and Bin Yang, "CMGAN: Conformer-based metric-GAN for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2477–2493, 2024.
- [22] Yi Luo and Nima Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [23] Fahimeh Bahmaninezhad, Jian Wu, Rongzhi Gu, Shi-Xiong Zhang, Yong Xu, Meng Yu, and Dong Yu, "A comprehensive study of speech separation: Spectrogram vs waveform separation," in *Proc. Interspeech*, 2019, pp. 4574–4578.
- [24] David Diaz-Guerra, Antonio Miguel, and Jose 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] Douglas B. Paul and Janet Baker, "The design for the Wall Street Journal-based CSR corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.
- [26] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015.
- [27] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey, "SDR – half-baked or well done?," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.
- [28] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. International Conference on Machine Learning (ICML)*, 2023.