Array Geometry-Robust Attention-Based Neural Beamformer for Moving Speakers

Marvin Tammen[2*], *Tsubasa Ochiai*[1], *Marc Delcroix*[1], *Tomohiro Nakatani*[1], *Shoko Araki*[1], *Simon Doclo*[2]

[1]NTT Corporation, Japan; [2]Carl von Ossietzky Universität Oldenburg, Germany

marvin.tammen@uol.de

Abstract

Although mask-based beamforming is a powerful speech enhancement approach, it often requires manual parameter tuning to handle moving speakers. Recently, this approach was augmented with an attention-based spatial covariance matrix aggregator (ASA) module, enabling accurate tracking of moving speakers without manual tuning. However, the deep neural network model used in this module is limited to specific microphone arrays, necessitating a different model for varying channel permutations, numbers, or geometries. To improve the robustness of the ASA module against such variations, in this paper we investigate three approaches: training with random channel configurations, employing the transform-average-concatenate method to process multi-channel input features, and utilizing robust input features. Our experiments on the CHiME-3 and DEMAND datasets show that these approaches enable the ASA-augmented beamformer to track moving speakers across different microphone arrays unseen in training.

Index Terms: multi-channel speech enhancement, moving speaker, mask-based beamformer, array geometry-robust processing

1. Introduction

In many speech communication applications, the microphone signals are corrupted by ambient noise, reducing speech quality and intelligibility as well as degrading the performance of automatic speech recognition (ASR) systems. When multiple microphones are available, good noise reduction performance with low speech distortion can be achieved using beamforming, provided that accurate estimates of the required spatial covariance matrices (SCMs) are available [1, 2].

In mask-based beamformers, the SCM estimation task has often been offloaded to deep neural networks (DNNs) [3–13]. These beamformers have demonstrated remarkable performance in recent ASR challenges such as the CHiME-4 challenge, while typically being applicable to arbitrary channel configurations, i.e., to arbitrary permutations and numbers of channels as well as associated microphone array geometries. However, most studies have focused on stationary acoustic scenarios, where the SCMs are estimated across entire utterances [3, 5, 7, 11]. This approach falls short in realistic acoustic scenarios involving moving speakers, where the SCMs are inherently time-varying. Various heuristic tracking methods, such as block-online estimation or recursive smoothing, have been proposed [6]. However, these methods heavily rely on manual tuning of parameters such as forgetting factors, which are highly dependent on the acoustic scenario, potentially leading to poor tracking performance.

To avoid such manual tuning and achieve a better tracking performance, a mask-based beamformer employing an attention-based SCM aggregator (ASA) module has been proposed in [13]. The ASA module temporally aggregates instantaneous estimates of the SCMs to compute time-varying speech and noise SCMs. In [13], it was demonstrated that the ASA module accurately tracks moving speakers, outperforming heuristic tracking methods. However, since the employed training procedure, DNN architecture, and input features depend on the channel configuration, the mask-based beamformer with ASA lost the ability to operate with arbitrary microphone array configurations, one of the key benefits of conventional mask-based beamformers.

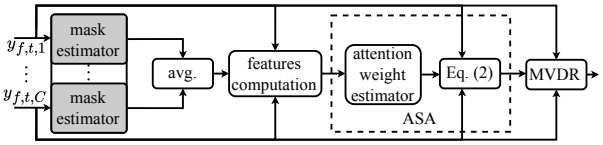


Figure 1: *Overview of mask-based MVDR beamformer with ASA. Grey vertically stacked boxes share weights.*

Aiming at realizing a mask-based beamformer with ASA for arbitrary microphone arrays, in this paper we propose three approaches extending the prior work in [13]. First, we investigate incorporating random channel configurations in the training procedure to prevent the DNN from overfitting to specific channel permutations and channel numbers. Second, we propose to employ the transform-average-concatenate (TAC) method [14] in the ASA module to process multi-channel features, allowing for any channel number and enabling permutation invariance. The TAC method was originally proposed for channel permutation-invariant multi-channel source separation and has been successfully employed, e.g., in time-frequency masking algorithms [15, 16] and stationary mask-based beamformers [11]. Third, we investigate utilizing input features that are less sensitive to variations of the channel configuration than the input features in [13]. Through experiments on the CHiME-3 [17] and DEMAND [18] datasets including moving speakers, we demonstrate the benefit of jointly integrating the three proposed approaches into the ASA module. Notably, our proposed approaches not only maintain high performance under matched conditions but also yield a good speech enhancement performance even for microphone arrays unseen during training, consistently outperforming a baseline mask-based beamformer with recursive smoothing and the mask-based beamformer with the original ASA in [13].

2. Mask-Based Beamformer With Attention-Based SCM Aggregator

In this section, we provide an overview of the mask-based beamformer with ASA [13], depicted in Fig. 1. The minimum variance distortionless response (MVDR) beamformer is described in Section 2.1, the estimation of the required SCMs and the time-frequency masks is described in Section 2.2, and the computation of the features and the ASA module are described in Section 2.3.

[*]This work was done during an internship at NTT Corporation.

2.1. MVDR Beamformer

We consider an acoustic scenario with a single moving speaker and additive noise in a reverberant room, recorded by a set of C microphones. In the short-time Fourier transform (STFT) domain, the vector comprising the C noisy microphone signals can be written as $\mathbf{y}_{f,t} = [y_{f,t,c=1}, ..., y_{f,t,c=C}]^{\mathsf{T}} \in \mathbb{C}^C$, where f, t, and c denote the frequency bin index, the time frame index, and the channel index, respectively, and $\cdot^{\mathsf{T}}$ denotes the transpose operator. Assuming that the (time-varying) acoustic transfer function between the speaker and the microphones is shorter than the STFT frame length, the noisy vector can be written as $\mathbf{y}_{f,t} = \mathbf{h}_{f,t} s_{f,t} + \mathbf{n}_{f,t}$, where $\mathbf{h}_{f,t} \in \mathbb{C}^C$, $s_{f,t} \in \mathbb{C}$, and $\mathbf{n}_{f,t} \in \mathbb{C}^C$ denote the acoustic transfer function, the speech source, and the additive noise component, respectively.

In beamforming approaches, the target speech component $x_{f,t,c=r} = h_{f,t,c=r} s_{f,t}$ at a reference microphone r is typically estimated by applying a linear filter $\mathbf{w}_{f,t} \in \mathbb{C}^C$ to the noisy vector, i.e., $\widehat{x}_{f,t,r} = \mathbf{w}_{f,t}^{\mathsf{H}} \mathbf{y}_{f,t}$, where $\cdot^{\mathsf{H}}$ denotes the conjugate transpose operator. Aiming at minimizing the output noise power spectral density while leaving the target speech component undistorted, the MVDR beamformer can be derived as [19]:

$$\mathbf{w}_{f,t} = \frac{\left(\mathbf{\Phi}_{f,t}^{n}\right)^{-1} \mathbf{\Phi}_{f,t}^{x}}{\mathrm{tr}\left(\left(\mathbf{\Phi}_{f,t}^{n}\right)^{-1} \mathbf{\Phi}_{f,t}^{x}\right)} \mathbf{u}_r, \tag{1}$$

where $\mathbf{\Phi}_{f,t}^{x} \in \mathbb{C}^{C \times C}$ and $\mathbf{\Phi}_{f,t}^{n} \in \mathbb{C}^{C \times C}$ denote the speech and noise SCMs, respectively, $\mathrm{tr}(\cdot)$ denotes the trace operator, and $\mathbf{u}_r \in \{0,1\}^C$ denotes a selection vector with a 1 as the r-th element and 0 otherwise.

2.2. Spatial Covariance Matrix Estimation

To implement the MVDR beamformer in (1), estimates of the speech and noise SCMs $\mathbf{\Phi}_{f,t}^{x}$ and $\mathbf{\Phi}_{f,t}^{n}$ are required. To allow estimating time-varying SCMs, in [13] the following temporal aggregation mechanism has been proposed:

$$\widehat{\mathbf{\Phi}}_{f,t}^{\nu} = \sum_{\tau=1}^{T} a_{t,\tau}^{\nu} \underbrace{m_{f,\tau}^{\nu} \mathbf{y}_{f,\tau} \mathbf{y}_{f,\tau}^{\mathsf{H}}}_{=\widehat{\mathbf{\Psi}}_{f,\tau}^{\nu}}, \tag{2}$$

where $\nu \in \{x, n\}$ indicates the speech or noise component, $m_{f,t}^{\nu}$ denotes a time-frequency mask, $\widehat{\mathbf{\Psi}}_{f,\tau}^{\nu} \in \mathbb{C}^{C \times C}$ is an instantaneous SCM (ISCM) estimate, and T denotes the number of time frames. The frequency-independent attention weights $\mathbf{a}_t^{\nu} = \left[a_{t,\tau=1}^{\nu}, ..., a_{t,\tau=T}^{\nu}\right]^{\mathsf{T}} \in \mathbb{R}^T$ control how the ISCM estimates are temporally aggregated to yield estimates of the speech and noise SCMs at time frame t. The time-frequency mask is typically obtained by applying a DNN-based mask estimator independently for each channel, followed by averaging across channels, i.e., $m_{f,t}^{\nu} = \frac{1}{C} \sum_{c=1}^{C} m_{f,t,c}^{\nu}$.

2.3. Attention Weight Estimation

To obtain the attention weights, a self-attention-based DNN (a transformer encoder [20]) is employed, i.e.,

$$\{\mathbf{a}_t\}_{t=1}^{T} = \mathrm{DNN}\left(\{\mathbf{i}_t\}_{t=1}^{T}; \mathbf{\Lambda}\right), \tag{3}$$

where $\mathbf{a}_t = [(\mathbf{a}_t^x)^{\mathsf{T}}, (\mathbf{a}_t^n)^{\mathsf{T}}]^{\mathsf{T}}$ denotes the attention weights at time frame t, $\mathbf{i}_t = [(\mathbf{i}_t^x)^{\mathsf{T}}, (\mathbf{i}_t^n)^{\mathsf{T}}]^{\mathsf{T}}$ denotes the input features at time frame t, and $\mathbf{\Lambda}$ denotes the parameters of the DNN[2]. As illustrated in Fig. 2 (top), the input features are first transformed into a time-varying embedding vector via a linear layer. This embedding vector then passes through several multi-head attention (MHA) encoder blocks, each

[2] In [13], $\mathrm{DNN}^x(\cdot)$ and $\mathrm{DNN}^n(\cdot)$ were used separately for the speech and noise components. Our preliminary experiments showed a similar or better performance at a lower computational complexity when using a single DNN.

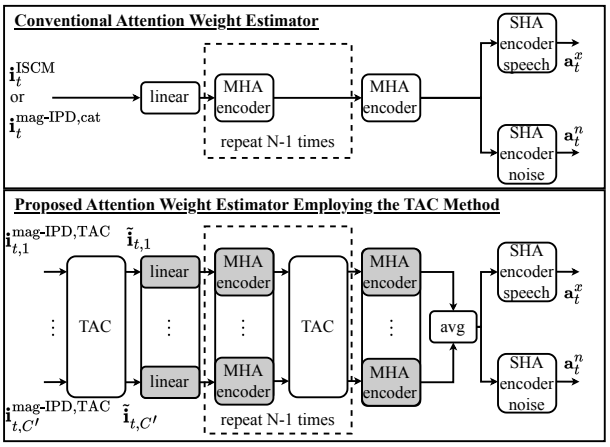


Figure 2: *Attention weight estimator employing different approaches to process multi-channel features. Grey vertically stacked boxes share weights.*

comprising multi-head self-attention layers and position-wise feedforward layers, which are all interconnected through residual connections. Finally, the attention weights $\mathbf{a}_t$ are extracted from two separate speech and noise single-head attention (SHA) layers. In [13], the ISCMs defined in (2) are used as the speech and noise input features, i.e.,

$$\mathbf{i}_{f,t}^{\nu,\text{ISCM}} = \left[\Re\left(\mathrm{vec}(\widehat{\mathbf{\Psi}}_{f,t}^{\nu})^{\mathsf{T}}\right), \Im\left(\mathrm{vec}(\widehat{\mathbf{\Psi}}_{f,t}^{\nu})^{\mathsf{T}}\right)\right]^{\mathsf{T}} \in \mathbb{R}^{2C^2}, \tag{4}$$

where $\mathrm{vec}(\cdot)$ denotes a reshaping of a $C \times C$-dimensional matrix into a vector of length C^2. The speech and noise features are concatenated along the frequency dimension, resulting in $\mathbf{i}_t^{\text{ISCM}} \in \mathbb{R}^{4FC^2}$.

As described in Section 2.2, the attention weights $\mathbf{a}_t$ control the temporal aggregation of the speech and noise ISCMs. Although the formulation in (2) allows for a potential application across different microphone arrays, it should be noted that the approach proposed in [13] does depend on the channel configuration. More specifically, the training procedure used a fixed channel configuration, not accounting for channel configuration variability, while the DNN architecture used a fixed input layer size, and the ISCM features in (4) simultaneously incorporate spatial and spectro-temporal information, making them sensitive to the channel configuration considered during training.

3. Proposed Approaches to Improve Robustness Against Channel Configuration Variations

In this section, we propose three approaches to improve the robustness of the mask-based beamformer with ASA against channel configuration variations.

3.1. Training With Random Channel Configurations

To prevent the DNN from overfitting to specific channel permutations, channel numbers, and microphone array geometries, a straightforward approach is to integrate random channel configurations into the training procedure. Assuming that a single microphone array with $C_{\max}$ channels is available for training, for each minibatch a channel number C' is drawn from the uniform random distribution $\mathcal{U}(2,\ C_{\max})$. From the available $C_{\max}$ channels, C' channels are then selected in random permutation, resulting in random microphone subarrays.

3.2. TAC Method to Process Multi-Channel Features

To accommodate a variable number of input channels in the training of the DNN with fixed input layer size, zero-padding up to $C_{\max}$ channels can be applied. However, this approach may sacrifice upper

bound speech enhancement performance for robustness, since the DNN needs to learn to deal with zero-padded input features, while also being limited to $C' \leq C_{\max}$ channels. To deal with this issue, we propose to employ the TAC method [14] to process multi-channel features in the attention weight estimator, as depicted in Fig. 2 (bottom). A TAC block takes as input a set of feature streams $\{\mathbf{z}_{t,c} \in \mathbb{R}^D\}_{c=1}^{C'}$ with variable C' and a channel-independent feature dimension D, shares information across the streams in a non-linearly transformed space, and outputs a set of modified feature streams $\{\tilde{\mathbf{z}}_{t,c} \in \mathbb{R}^D\}_{c=1}^{C'}$. We adopt the efficient TAC implementation from [15], obtaining the modified feature stream at time frame t and channel c as:

$$\tilde{\mathbf{z}}_{t,c} = \left[\mathrm{ReLU}(\mathbf{L}_1 \mathbf{z}_{t,c})^\mathsf{T}, \frac{1}{C'} \sum_{\mu=1}^{C'} \mathrm{ReLU}(\mathbf{L}_2 \mathbf{z}_{t,\mu})^\mathsf{T} \right]^\mathsf{T}, \quad (5)$$

where $\mathbf{L}_1 \in \mathbb{R}^{(D/2) \times D}$ and $\mathbf{L}_2 \in \mathbb{R}^{(D/2) \times D}$ denote trainable linear transforms shared across all channels. The modified feature streams contain channel-specific information as well as information affected by all channels in a permutation-invariant fashion due to the combination of weight sharing and the application of the permutation-invariant averaging operation.

The TAC method is integrated into the attention weight estimator by interleaving TAC blocks with C' parallel MHA encoder blocks sharing the same parameters (see Fig. 2, bottom). After N stacks of parallel interleaved TAC blocks and MHA encoder blocks, the streams are averaged and passed to the final SHA speech and noise encoder blocks. This integration enables handling a varying channel number C' (even $C' > C_{\max}$) and ensures invariance to the channel permutation. This significantly enhances the flexibility and applicability of the attention weight estimator across diverse channel configurations, without necessitating modifications in the DNN architecture or hyperparameters.

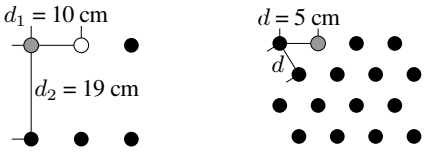
3.3. Input Features

Due to the definition of the ISCM in (2), the input features in (4) simultaneously encode inter-microphone level differences (ILDs) as well as inter-microphone phase differences (IPDs) and hence strongly depend on the microphone array geometry. In addition, the $4FC^2$-dimensional features $\mathbf{i}_t^{\mathrm{ISCM}}$ below (4) are incompatible with the TAC method, since it requires channel-wise feature streams with a channel number-independent feature dimension.

To address these issues, we propose to adopt alternative channel-wise feature streams (denoted as mag-IPD features), defined as:

$$\mathbf{i}_{f,t,c}^{\nu,\text{mag-IPD}} = \left[|\widehat{\nu}_{f,t,c}|^2, \cos\left(\widehat{\delta}_{f,t,c}\right), \sin\left(\widehat{\delta}_{f,t,c}\right) \right]^\mathsf{T}, \quad (6)$$

where $\widehat{\delta}_{f,t,c} = \angle \widehat{\nu}_{f,t,c} - \angle \widehat{\sigma}_{f,t}$ denotes the difference between the unwrapped phases of the masked STFT coefficients $\widehat{\nu}_{f,t,c} = m_{f,t}^{\nu} y_{f,t,c}$ and the channel-averaged masked STFT coefficients $\widehat{\sigma}_{f,t} = \frac{1}{C} \sum_{c=1}^{C} \widehat{\nu}_{f,t,c}$, and cos and sin have been applied to result in a smooth phase representation. The features proposed in (6) differ from those in [15], which used $|\widehat{\sigma}_{f,t}|^2$ instead of $|\widehat{\nu}_{f,t,c}|^2$ and the phase component instead of its cosine and sine, yielding a worse performance in our preliminary experiments. Concatenating the speech and noise features along the frequency dimension yields C streams of $6F$-dimensional features $\mathbf{i}_{t,c}^{\text{mag-IPD,TAC}}$ (see Fig. 2, bottom). We hypothesize that these features are less sensitive to the channel configuration than the features in (4) because they do not explicitly depend on channel pairs and they effectively separate the channel configuration-dependent IPD information from magnitude information, which is less influenced by the channel configuration. In addition, we employ the proposed features with the conventional attention weight estimator, in which case we concatenate the speech and noise features along the frequency and channel dimensions, yielding $6FC$-dimensional features $\mathbf{i}_t^{\text{mag-IPD,cat}}$ (see Fig. 2, top).



(a) *CHiME-3 [17]* (b) *DEMAND [18]*

Figure 3: *Considered microphone array geometries. Grey circles denote the reference and white circles denote unused microphones.*

4. Experiments

4.1. Datasets

To evaluate the effectiveness of the proposed approaches, we constructed datasets of simulated moving speakers in noisy conditions using speech signals from the Wall Street Journal (WSJ0) corpus [21] and noise recordings from the CHiME-3 [17] and DEMAND [18] corpora. We constructed two datasets with different microphone array geometries (illustrated in Fig. 3), both with a sampling frequency of 16 kHz. Similarly as in [13], we simulated speakers moving on a linear trajectory with constant speed using the gpuRIR tool [22] by generating room impulse responses (RIRs) at 128 positions on a line, with room width and depth uniformly drawn from the set of $\{3.0\,\mathrm{m}, 3.5\,\mathrm{m}, 4.0\,\mathrm{m}, 4.5\,\mathrm{m}, 5.0\,\mathrm{m}\}$, room height equal to 2.5 m, reverberation time T_{60} drawn uniformly between 0.1 s and 0.3 s, and the microphone array randomly placed in the room. We added the speech signal convolved with the simulated RIRs and the recorded noise signals at signal-to-noise-ratios (SNRs) between 2 dB and 8 dB.

The first dataset consists of simulated utterances based on the WSJ0 speech and CHiME-3 noise signals, resulting in a maximum number of $C_{\max} = 5$ channels available for training (excluding the rear-facing second channel). This dataset was used for training, development, and evaluation. The second dataset consists of simulated utterances based on the WSJ0 speech and DEMAND noise signals, resulting in 16 available channels. This dataset was only used for evaluation.

During evaluation, we considered a matched condition and several mismatched conditions. "matched" represents the CHiME-3-based evaluation dataset with a fixed channel permutation and the channel number $C' = C_{\max} = 5$, similar to the fixed training condition. To evaluate a mismatch in terms of the channel permutation, we randomly permuted the channels from the CHiME-3-based evaluation dataset. To evaluate a mismatch in terms of the channel number, we selected the first $C' = 3$ channels from the CHiME-3-based evaluation dataset. To evaluate a mismatch in terms of the microphone array geometry, we randomly selected $C' = 5$ channels from the DEMAND-based evaluation dataset. This procedure allows for diverse microphone array geometries, e.g., including linear, triangular, rectangular, and trapezoidal shapes, some of which are not realizable with the CHiME-3 microphone array used for training (see Fig. 3). To evaluate a mismatch in terms of both the channel number and the microphone array geometry, we randomly selected $C' = 3$ channels from the DEMAND-based evaluation dataset. In all evaluation conditions, the reference channel was chosen as depicted in Fig. 3. We created 30 000, 2000, and 2000 noisy utterances for training, development, and each evaluation dataset, respectively.

4.2. Settings

We mostly followed the experimental settings presented in [13] to increase comparability with the associated results. We trained the attention weight estimator in an end-to-end manner, utilizing the scale-dependent SNR loss function [23] at the output of the mask-based beamformer (see Fig. 1), with the reverberant clean speech component at the reference microphone as the target signal. During training, we

Table 1: *Mean PESQ and SDR values for the noisy mixtures, a mask-based MVDR beamformer with recursive smoothing using a fixed forgetting factor, and the mask-based MVDR beamformer with ASA employing different attention weight estimators, evaluated on datasets corresponding to a matched condition and various mismatched conditions.*

					matched		mismatched in terms of							
							Permutation		Number		Geometry		Number & Geom.	
		config.	features	use TAC	PESQ	SDR	PESQ	SDR	PESQ	SDR	PESQ	SDR	PESQ	SDR
1	mixture	—	—	—	1.37	5.19	1.37	5.19	1.37	5.19	1.38	3.12	1.38	3.12
2	recursive MVDR	—	—	—	2.04	10.18	2.04	10.18	1.73	9.05	2.00	8.94	1.73	7.40
3	baseline [13]	fixed	ISCM	False	**2.64**	16.34	2.31	13.72	1.84	10.66	2.19	11.32	1.71	7.39
4	proposed	fixed	mag-IPD	False	2.57	16.27	2.40	14.70	1.84	10.71	2.15	11.01	1.77	8.34
5	proposed	fixed	mag-IPD	True	2.62	**16.39**	**2.62**	**16.39**	2.05	12.55	2.20	11.84	1.87	9.25
6	proposed	random	ISCM	False	2.42	14.37	2.42	14.36	1.96	11.86	2.18	11.55	1.93	10.02
7	proposed	random	mag-IPD	False	2.53	15.85	2.52	15.86	2.32	14.07	2.18	11.82	1.99	10.54
8	proposed	random	mag-IPD	True	2.59	16.02	2.59	16.02	**2.34**	**14.15**	**2.21**	**12.35**	**2.03**	**11.34**

used oracle Wiener-like time-frequency masks [24] to compute the ISCMs in (2) and optimized only the trainable parameters of the attention weight estimator. During the evaluation, we used a time-frequency mask estimator based on a temporal convolutional network architecture [25]. For the attention weight and time-frequency mask estimators, we adopted the DNN and training hyperparameters in [13], except for using a single DNN for both the speech and noise components (see Section 2.3). For the TAC blocks, we adopted the implementation proposed in [15], consisting of linear layers and ReLU activations (cf. (5)).

In addition to the mask-based MVDR beamformer with the original ASA in [13], we considered a mask-based MVDR beamformer with recursive smoothing using a fixed (frequency-independent) forgetting factor that corresponds to a time constant of $1.6\,\mathrm{s}$ (tuned according to the highest signal-to-distortion-ratio (SDR) values under the matched evaluation condition) as a baseline algorithm [4, 13]. For the STFT, we used a Hann window with a frame length of $64\,\mathrm{ms}$ and $16\,\mathrm{ms}$ shift.

We evaluated the speech enhancement performance in terms of perceptual evaluation of speech quality (PESQ) [26] and SDR [27] (allowing for distortions caused by time-invariant filters), with the reverberant clean speech component at the reference microphone as the reference signal.

4.3. Results

Table 1 shows the mean PESQ and SDR values for the noisy mixtures, the mask-based beamformer with recursive smoothing (described in the previous section), and for the mask-based beamformer with ASA employing different attention weight estimators (baseline estimator in [13] and proposed estimators). In this table, "config" indicates whether the channel permutation and number were fixed or randomized during training (see Section 3.1); "features" represents the utilized input features, either the ISCM features in (4) or the proposed mag-IPD features in (6); "use TAC" indicates whether TAC was employed or not. We evaluated these beamformers both under a matched condition as well as under various mismatched conditions described in Section 4.1.

The results in Table 1 show that under all conditions both the mask-based beamformer with recursive smoothing as well as the mask-based beamformer with ASA (for all attention weight estimators) substantially improve the PESQ and SDR values compared to the noisy mixtures. Under the matched condition, it can be observed that models trained with a fixed channel configuration (rows 3-5) achieve the highest PESQ and SDR values. This is expected as these models can exploit the specific spatial information seen during training, representing an upper bound in performance.

Under mismatched conditions, the baseline model (row 3) shows notable performance degradation, particularly in terms of channel number and microphone array geometry. The model employing mag-IPD features (row 4) exhibits a similar performance as the baseline model in most conditions, except for a reduced performance drop under the channel permutation mismatch. The model employing ISCM features with randomized training configurations (row 6) demonstrates similar robustness across mismatched conditions as the model in row 4, albeit with a worse performance under the matched condition, highlighting a trade-off between robustness and upper bound performance. The incorporation of mag-IPD features and the TAC method (row 5) further mitigates performance drops across all mismatch conditions, completely alleviating the drop under the channel permutation mismatch while maintaining strong matched condition performance. The model combining mag-IPD features, TAC, and randomized training configurations (row 8) achieves the most consistent high performance, performing similarly as the best model under the matched condition and the channel permutation mismatch conditions (row 5), as well as outperforming all models under channel number and microphone array geometry mismatches. The results clearly show that the combination of training with random channel configurations, employing the TAC method, and using the mag-IPD-based input features resulted in a significantly higher speech enhancement performance compared to the baseline model [13] (significance determined using a two-sided T-test with Bonferroni correction).

It should be emphasized that the evaluation included diverse microphone array geometries by randomly selecting channels from the DEMAND-based evaluation dataset, i.e., "Geometry" and "Number & Geom." in Table 1. Hence, the results show that the mask-based beamformer with ASA using the combination of all proposed approaches can perform noise reduction for moving speakers and arbitrary microphone arrays, consistently outperforming the mask-based beamformer with recursive smoothing and the baseline mask-based beamformer with the original ASA.

5. Conclusion

In this paper, we proposed several approaches to improve the robustness of the mask-based beamformer with ASA against channel configuration variations. These approaches include the integration of random channel configurations during training, employing the TAC method to process multi-channel features (allowing for any channel number and enabling permutation invariance), as well as using mag-IPD features that are robust against channel configuration variations. Experiments using the CHiME-3 and DEMAND datasets suggest that the mask-based beamformer with ASA integrating the proposed approaches can perform noise reduction for moving speakers and arbitrary microphone arrays. Future research will extend this investigation to explore more diverse channel configurations during training and evaluation as well as address the computational complexity of the proposed TAC integration.

6. References

[1] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel Signal Enhancement Algorithms for Assisted Listening Devices: Exploiting Spatial Diversity Using Multiple Microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[2] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

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

[4] T. Higuchi, N. Ito, T. Yoshioka, and T. 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)*, Shanghai, China, Mar. 2016, pp. 5210–5214.

[5] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. L. Roux, "Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 1981–1985.

[6] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring Practical Aspects of Neural Mask-Based Beamforming for Far-Field Speech Recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 6697–6701.

[7] Z.-Q. Wang, P. Wang, and D. Wang, "Multi-microphone Complex Spectral Mapping for Utterance-wise and Continuous Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2001–2014, 2021.

[8] J. Casebeer, J. Donley, D. Wong, B. Xu, and A. Kumar, "NICE-Beam: Neural Integrated Covariance Estimators for Time-Varying Beamformers," *arXiv:2112.04613*, pp. 1–5, Dec. 2021.

[9] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "ADL-MVDR: All Deep Learning MVDR Beamformer for Target Speech Separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 6089–6093.

[10] Y. Wang, A. Politis, and T. Virtanen, "Attention-Driven Multichannel Speech Enhancement in Moving Sound Source Scenarios," *arXiv:2312.10756*, pp. 1–5, Dec. 2023.

[11] A. Jukić, J. Balam, and B. Ginsburg, "Flexible Multichannel Speech Enhancement for Noise-Robust Frontend," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2023, pp. 1–5.

[12] M. Tammen and S. Doclo, "Parameter Estimation Procedures for Deep Multi-Frame MVDR Filtering for Single-Microphone Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1–13, Aug. 2023.

[13] T. Ochiai, M. Delcroix, T. Nakatani, and S. Araki, "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.

[14] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end Microphone Permutation and Number Invariant Multi-channel Speech Separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 6394–6398.

[15] T. Yoshioka, X. Wang, D. Wang, M. Tang, Z. Zhu, Z. Chen, and N. Kanda, "VarArray: Array-Geometry-Agnostic Continuous Speech Separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 6027–6031.

[16] D. Wang, Z. Chen, and T. Yoshioka, "Neural Speech Separation Using Spatially Distributed Microphones," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 339–343.

[17] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, USA, Dec. 2015, pp. 504–511.

[18] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. Meetings on Acoustics (ICA)*, Montreal, Canada, 2013, pp. 1–6.

[19] M. Souden, J. Benesty, and S. 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, Feb. 2010.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Proc. Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Long Beach, USA, Dec. 2017, pp. 5998–6008.

[21] D. B. Paul and J. M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," in *Proc. Workshop on Speech and Natural Language*, New York, USA, Feb. 1992, pp. 357–362.

[22] 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, Feb. 2021.

[23] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 626–630.

[24] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-Sensitive and Recognition-Boosted Speech Separation Using Deep Recurrent Neural Networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Queensland, Australia, Apr. 2015, pp. 708–712.

[25] Y. Luo and N. 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, Aug. 2019.

[26] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ) - A New Method for Speech Quality Assessment of Telephone Networks and Codecs," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, USA, May 2001, pp. 749–752.

[27] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.