Mask-Based Neural Beamforming for Moving Speakers With Self-Attention-Based Tracking

Tsubasa Ochiai, Member, IEEE, Marc Delcroix, Senior Member, IEEE, Tomohiro Nakatani, Fellow, IEEE, and Shoko Araki, Fellow, IEEE

Abstract—Beamforming is a powerful tool designed to enhance speech signals from the direction of a target source. Computing the beamforming filter requires estimating spatial covariance matrices (SCMs) of the source and noise signals. Time-frequency masks are often used to compute these SCMs. Most studies of mask-based beamforming have assumed that the sources do not move. However, sources often move in practice, which causes performance degradation. In this paper, we address the problem of mask-based beamforming for moving sources. We first review classical approaches to tracking a moving source, which perform online or blockwise computation of the SCMs. We show that these approaches can be interpreted as computing a sum of instantaneous SCMs weighted by attention weights. These weights indicate which time frames of the signal to consider in the SCM computation. Online or blockwise computation assumes a heuristic and deterministic way of computing these attention weights that, although simple, may not result in optimal performance. We thus introduce a learning-based framework that computes optimal attention weights for beamforming. We achieve this using a neural network implemented with self-attention layers. We show experimentally that our proposed framework can greatly improve beamforming performance in moving source situations while maintaining high performance in non-moving situations, thus enabling the development of mask-based beamformers robust to source movements.

Index Terms—Array processing, mask-based neural beamformer, moving source, self-attention network, time-varying filter.

I. INTRODUCTION

MICROPHONE array signal processing [1], [2], [3], which uses spatio-temporal information obtained with multiple microphones, has been an active research field for several decades and plays an important role in the development of many applications. In particular, multichannel linear filtering using a microphone array, i.e., beamforming, has been used extensively to design speech enhancement systems for hearing aids [4], [5] and for noise-robust automatic speech recognition (ASR) systems [6], [7], [8]. Recently, the mask-based beamforming approaches [9], [10], [11] have attracted increased attention because they were shown to be particularly effective in reducing noise or the effect of interference speakers in recent robust ASR challenges [12], [13].

A beamformer exploits the spatial information about the target and interfering sources derived from spatial covariance matrices (SCMs) to emphasize the signals coming from a target source direction while suppressing the interfering signals. The mask-based beamformer exploits time-frequency masks derived from neural networks (NNs) [9], [10] or other source models such as complex Gaussian mixture models (cGMMs) [11] to compute the SCMs. SCMs capture the spatial information and are thus sensitive to source movements. Most studies involving mask-based beamformers avoided this issue by assuming that the target and interfering sources do not move within an utterance. However, this hypothesis may not hold in general, especially when considering more realistic situations such as sound captured by a smart speaker or robots, where the target speaker or interference speakers could, for example, walk around the room while talking. In this paper, we address the problem of designing a mask-based beamformer that is robust to moving sources by proposing a novel estimation framework of the beamforming filters that can track the source movements.

Mask-based beamformers compute the source and interference SCMs by averaging over time the outer product of the multi-channel observation vectors (i.e., the vector of the multi-channel observed signal at each time-frequency bin) masked with the time-frequency masks. We can compute time-invariant SCMs over an entire utterance if we assume that the sources are not moving. This procedure results in a time-invariant beamformer.

Adapting this framework to a moving source scenario requires estimation of time-varying SCMs and beamforming filters, which reflect changes in the acoustic conditions, i.e., the source positions. We can estimate the time-varying SCMs using online or blockwise processing. For example, the online mask-based beamformer [7], [11], [14], [15] sequentially updates the SCMs. These approaches estimate one SCM and the resultant beamforming filters for each frame or block, not for the entire utterance, and thus they could potentially deal with moving sources. However, they require tuning hyperparameters, such as the forgetting factor and block size that may vary depending on, e.g., the speed of the sources. Consequently, these approaches may not track a source in an optimal way.

We can view the computing of SCMs by online or blockwise processing as limiting the range of frames that contribute to estimating the SCMs for each frame or block. That is, such processing replaces the averaging operation in the SCM
Time-varying mask-based beamformers have also been investigated using an NN to predict the time-varying SCMs computed with the attention weights. With this fully supervised scheme, we can learn to predict optimal attention weights that allow the beamforming to steer its directivity toward the position of the moving source for each frame, i.e., that enables implicit source tracking by the attention mechanism.

Note that time-varying beamformers are often investigated for online (sequential) systems that target processing with low latency, but they can also be used for offline systems to estimate better SCMs and beamforming filters such as those in a previous work [19]. Similarly, in this paper, we focus on offline processing that utilizes all of the information within an utterance. We could easily extend the proposed framework to sequential processing by restricting the use of future frames. We also evaluated the potential of our approach for sequential and latency-controlled processing in the experiment.

We tested the effectiveness of the proposed framework on moving source signals simulated using the Wall Street Journal (WSJ0) corpus [20] for the speech signals, dynamic room impulse responses computed with the gpuRIR toolkit [21], and background noise derived from CHiME-3 corpus [12]. Experimental results show that the proposed framework achieves better speech enhancement and ASR performance, i.e., signal-to-distortion ratio (SDR), perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), and word error rate (WER), compared to the conventional time-invariant, online, and blockwise beamforming frameworks. In addition, we confirmed that our proposed scheme could track a moving source by visualizing the directivity characteristics (i.e., beam patterns) of the time-varying beamformer computed with our proposed scheme.

The main contributions of this paper are as follows:

1) We provide a generalization of the time-varying SCM computation approaches and propose a learning-based (fully supervised) scheme to allow the design of time-varying mask-based beamformers that can track moving sources.

2) We introduce an self-attention-based NN that predicts the time frames that are relevant for computing the SCMs at a given time.

3) We design an experiment using simulated moving sources to compare the different approaches for tackling moving sources and show the superiority of our proposed framework for both speech enhancement and ASR.

The remainder of this paper is summarized as follows. In Section II, we briefly discuss prior works related to our approach. Section III describes the conventional mask-based beamforming framework. In Section IV, we first generalize the online and blockwise framework and then introduce the proposed time-varying beamforming framework with the attention weight estimation model. In Section V, we detail the experimental conditions of the moving source scenario and demonstrate the effectiveness of the proposed framework. Finally, we conclude this paper in Section VI.

II. RELATED WORKS

Here, we briefly review related speech enhancement approaches that deal with source movements.

A. Beamformer-Based Approach

1) Mask-Based Beamformer: A mask-based beamformer first computes a time-frequency mask, which indicates the time-frequency bins where the target source is dominant. The mask is used to compute the SCMs of the target source and noise, which are required to compute the beamformer coefficients.

There are currently two main research directions toward estimating the time-frequency masks for mask-based beamformers, i.e., spatial clustering [11] and NNs [9], [10]. The spatial clustering-based approaches estimate the time-frequency masks based on the spatial information, which is derived from the microphone array signals, and thus the estimation accuracy is affected by the movements of the source signals. On the other hand, the NN-based approaches estimate the time-frequency masks mainly based on the spectral information, which can be derived even from a single microphone signal, and thus, in principle, these methods are not affected by source movements. Therefore, we adopt the NN-based approach to estimate the time-frequency masks of moving sources.

Many related studies have investigated online/low-latency processing for mask-based beamformers, e.g., [7], [11], [14], [15]. Most of these studies focused on the online computation of the beamformer coefficients given the masks. However, only a few approaches have actually been evaluated with moving source scenarios. For example, in a prior work [14], the authors introduced the block-online processing of a mask-based beamformer to deal with a moving source scenario.

Other work [15] investigated using an NN to predict the forgetting factor for online computation of the SCM, but it was not evaluated on moving source scenarios. Our approach can be considered the generalization of that previous effort [15], where we extend the formalization to offline processing and introduce self-attention-based NNs that naturally generalize the computation of the time-varying SCMs of conventional online and blockwise approaches. Furthermore, we evaluated and analyzed the behavior of the proposed approach on a moving source dataset.

Time-varying mask-based beamformers have also been investigated to improve performance for offline processing. For example, our previous effort assumed a time-varying noise
conduct source localization, which avoids the impact of localization errors on beamforming performance.

From a different perspective, the conventional source localization-based beamformer constructs the steering vector of source signals given the direction of arrival (DOA) information, by following the plane wave assumption [27], [33]. However, such an assumption would not always hold in complicated acoustic conditions. For example, one study [34] reported that a gain and delay steering vector achieves better speech enhancement performance compared to a delay-only (DOA-based) steering vector. The mask-based beamformer has been proposed for estimating a better steering vector (or SCM for source signals) [11], [35], and recently it has been widely used as the beamforming framework (e.g., in CHiME challenges [12], [13]) instead of a DOA-based beamformer. Our proposed framework is based on the mask-based beamforming framework, which would avoid the impact of the assumption mismatch on beamforming performance.

C. Blind Source Separation Approach

Source movements are also a problem for microphone-array-based blind source separation. Recently, several studies proposed estimating time-invariant separation filters that are robust to source movements [36], [37], [38]. However, these approaches may deal with only relatively small source movements because the filters are time-invariant.

In contrast, our proposed framework estimates the frame-by-frame time-varying filters that can track a source even for large movements such as 360° movements in the experiments of Section V-E2.

III. CONVENTIONAL MASK-BASED BEAMFORMER

A. Problem Definition

Let $Y_{t,f} = \{Y_{t,f,c=1}, \ldots, Y_{t,f,c=C}\} \in \mathbb{C}^C$ be a vector comprising the $C$-channel short-time Fourier transform (STFT) coefficients of the observed noisy signal at a time-frequency bin $(t, f)$, where $Y_{t,f,c} \in \mathbb{C}$ is the STFT coefficient for the $c$-th channel. Let $T$ and $F$ be the number of time frames and frequency bins, respectively. Assuming that the acoustic condition (i.e., the transfer function) is static within a short-time duration (i.e., a short time frame), the observed signal $Y_{t,f} \in \mathbb{C}^C$ can be approximately modeled as:

$$Y_{t,f} = H_{t,f}S_{t,f} + N_{t,f},$$

(1)

where $S_{t,f} \in \mathbb{C}$ and $N_{t,f} \in \mathbb{C}^C$ denote the speech source and additive noise signals at the time-frequency bin $(t, f)$, respectively. $H_{t,f} \in \mathbb{C}^C$ denotes the time-varying transfer function between the speech source and the microphones at a time-frequency bin $(t, f)$.

When the source is not moving, we can assume that the transfer function is static within an utterance, i.e., $H_{t,f} = H_{t}$, and thus use time-invariant beamformers to enhance the noisy speech signals. This is the scheme used in many studies and challenges [12], [13]. However, in general, the transfer function dynamically changes due to, e.g., the movements of the source, which is the situation we tackle in this paper. Therefore, we
assume the observation model of (1) and investigate the design of time-varying (frame-by-frame) mask-based beamformers to enhance the speech source $S_{t,f}$.

B. Minimum Variance Distortionless Response Beamformer

Given the observed noisy signal $Y_{t,f}$, a frequency-domain beamformer estimates the STFT coefficient of the enhanced speech, $\hat{S}_{t,f} \in \mathbb{C}$, as follows:

$$\hat{S}_{t,f} = w_{t,f}^H Y_{t,f},$$

(2)

where $w_{t,f} \in \mathbb{C}^C$ denotes a vector comprising the beamforming filter coefficients and $\hat{S}$ represents the conjugate transpose. We then obtain the time-domain enhanced signal, $\hat{s} \in \mathbb{R}^T$, by applying the inverse STFT to $\hat{S}_{t,f}$ and the overlapping add method, where $T$ denotes the duration of the time-domain signal.

We can compute the beamforming filter coefficients from the SCMs of the speech and noise signals. We adopt in this paper a widely used MVDR formalization, which computes the beamforming filter coefficients $w_{t,f}$ as follows [39]:

$$w_{t,f} = \left(\Phi_{t,f}^N\right)^{-1} \Phi_{t,f}^S u,$$

(3)

where $\Phi_{t,f}^S \in \mathbb{C}^{C \times C}$ and $\Phi_{t,f}^N \in \mathbb{C}^{C \times C}$ are the SCMs of the speech and noise signals at time-frequency bin $(t, f)$, respectively. $u \in \mathbb{R}^C$ is a one-hot vector representing the index of the reference microphone.

C. Mask-Based Spatial Covariance Matrix Estimation

The mask-based beamforming scheme relies on the sparseness property of speech signals in the STFT domain [40] to estimate the SCMs using time-frequency masks [9], [10], [11], [41], [42]. Here, the masks indicate the time-frequency bins where the source or noise is dominant. In the following, we briefly overview several commonly used options for estimating the SCMs from the time-frequency masks.

1) Time-Invariant SCM Computation: Assuming that the transfer function is static within the utterance, we can compute the time-invariant SCMs $\Phi^\nu_j$ as [9], [10], [11]:

$$\Phi^\nu_j = \sum_{\tau=t-L}^{t} \sum_{\nu' = 1}^{T} \frac{1}{m_{\tau,f}^{\nu'}} Y_{\tau,f}^\nu Y_{\tau,f}^\nu \hat{\Psi}_{\tau,f}^\nu,$$

(4)

where $m_{\tau,f}^{\nu'} \in [0,1]$ is a time-frequency mask and $\nu \in \{S, N\}$ are the indexes for speech and noise, respectively. By abuse of terminology, we call $\Psi_{t,f}$ the instantaneous SCM (ISCM) at time-frequency bin $(t, f)$. Because the SCMs are time-invariant, the beamforming filter coefficients computed with (3) are also time-invariant. Therefore, this approach cannot handle moving sources well.

2) Online SCM Computation: A conventional way to compute a time-varying SCM $\Phi^\nu_{t,f}$ is to use a recursive approach [7], [11], [14], [15]:

$$\Phi^\nu_{t,f} = \alpha \Phi^\nu_{t-1,f} + \Psi^\nu_{t,f}$$

(5)

$$= \sum_{\tau=t}^{T} \alpha^{t-\tau} \Psi_{\tau,f}^\nu,$$

(6)

where $\alpha$ denotes the forgetting factor, which gives exponentially less weight to the older ISCMs. With this approach, the SCMs and the beamforming filter coefficients are estimated at each time frame, which would allow tracking a source. However, the tracking speed depends on the forgetting factor. It may thus be challenging to tune this parameter to offer optimal performance for various conditions of source movement.

In this paper, we adopt the frame-by-frame update of the beamforming filters for the online processing to allow precise tracking instantaneously.

3) Blockwise SCM Computation: An alternative way of computing time-varying SCMs is to use blockwise processing [19], i.e., dividing a signal into consecutive time blocks and computing the SCMs for each block as follows:

$$\Phi^\nu_{t,f} = \sum_{\tau=t-L}^{t} \frac{1}{m_{\tau,f}^{\nu'}} \Psi_{\tau,f}^\nu,$$

(7)

where $L$ is a block size parameter that denotes the half span of the blocks, and thus $2L + 1$ frames are used for the SCM computation of each block.

Setting the block size requires a trade-off between using a large block size to allow computing reliable statistics and a small block size to allow better tracking. Therefore, as with the online SCM computation, tuning this parameter may be challenging and lead to sub-optimal performance.

In addition to the block size, we can consider the block shift, which determines how often we compute the SCMs and beamforming filter coefficients. In the experiments of Section V, we use a block shift of one frame, which means that the SCMs and the beamforming filter coefficients are computed for each frame like the online SCM computation described in Section III-C2.

IV. PROPOSED TIME-VARYING SCM COMPUTATION WITH SELF-ATTENTION-BASED WEIGHTING

A. Generalized Formulation of SCM Computation

We can express the different SCM computation approaches using a general formulation as:

$$\Phi^\nu_{t,f} = \sum_{\nu' = 1}^{T} c_{t\nu'\nu} \Psi^\nu_{t,f},$$

(8)

where $c_{t\nu'\nu} = \{c_{t\nu'\nu} = 1, \ldots, c_{t\nu'\nu} = T\} \in \mathbb{R}^T$ are weight coefficients that control the range for accumulating the statistics used to compute the SCMs for the $t$-th frame. The weight coefficients $c_{t\nu'\nu}$ determine which time frames to focus on when computing the SCMs at a given time frame (i.e., $t$) among all time frames (i.e., $t' = 1 \sim T$). In this paper, we refer to these weight coefficients as attention weights.

We can easily see that the SCM computation approaches discussed in Section III-C are special cases of the general formulation of (8). The time-invariant computation of (4) corresponds to:

$$c_{t't'} = \frac{1}{\sum_{\nu' = 1}^{T} m_{\tau'}^{\nu'}}$$

(9)
the online computation of (6) to,
\[ c^{t'}_{t,t'} = \begin{cases} \alpha^{t-t'} & t' \leq t, \\ 0 & t' > t, \end{cases} \] (10)
and the blockwise computation of (7) to setting,
\[ c^{t'}_{t,t'} = \begin{cases} \frac{1}{\sum_{t'=t-L}^{t+L} m^{t'}_{t,f}} & t' \in [t-L, \ldots, t+L], \\ 0 & t' \notin [t-L, \ldots, t+L]. \end{cases} \] (11)

The major contribution of this paper is to provide a generalization of the time-varying SCM computation approaches as shown in (8). This generalization allows us to derive a learning-based approach for computing the SCMs that are optimal for each time step by exploiting all frames of the signal.

B. Self-Attention-Based Time-Varying Attention Weight Estimation

1) Overall Procedure of Time-Varying Attention Weight Estimation: As mentioned above, the online and blockwise SCM computations use simple rules to compute the attention weights. These approaches would allow handling moving source scenarios, but such simple rules may not be necessarily optimal for tracking moving sources. In this section, we propose instead to design an NN to estimate optimal attention weights.

Fig. 1 illustrates the proposed estimation procedure of the time-varying SCMs with attention weight estimation neural network.

First, to make the ISCMs suitable for the NN’s input, we convert the ISCMs of all frequency bins at a given time frame \( t \) \( \{ \Psi_{t,f}^{\nu} \}_{f=1}^{F} \in \mathbb{C}^{F \times C \times C} \) into a real-valued vector \( \psi^{\nu}_{t} \in \mathbb{R}^{2FC^2} \) as:
\[ \psi^{\nu}_{t} = \text{Vectorize}(\{ \Psi_{t,f}^{\nu} \}_{f=1}^{F}), \] (12)
where \text{Vectorize}(\cdot) represents the unfolding operation that converts the complex-valued tensor \( \{ \Psi_{t,f}^{\nu} \}_{f=1}^{F} \) into the

real-valued vector \( \psi^{\nu}_{t} \), which contains the real and imaginary parts of all elements of the tensor.

We use the sequence of vectorized ISCMs \( \{ \psi^{\nu}_{t} \}_{t=1}^{T} \) as input to an NN that estimates the time-varying attention weight coefficients \( \{ c^{t'}_{t,t'} \}_{t=1}^{T} \) as follows:
\[ \{ c^{t'}_{t,t'} \}_{t=1}^{T} = \text{NN}^{\nu}(\{ \psi^{\nu}_{t} \}_{t=1}^{T}; \Lambda^{\nu}), \] (13)
where \( \text{NN}^{\nu}(\cdot) \) is the non-linear transformation of an NN and \( \Lambda^{\nu} \) denotes the learnable parameters of \( \text{NN}^{\nu}(\cdot) \). \( \text{NN}^{\nu}(\cdot) \) should predict attention weights that allow accumulating ISCMs from a similar direction to estimate reliable SCMs while making it possible to track a moving source. Since the input ISCMs capture information about the source direction, this behavior can naturally be implemented using an architecture for \( \text{NN}^{\nu}(\cdot) \) inspired by self-attention network [16], which estimates the weight coefficients focusing on the similarity between the input frames.

Fig. 2 summarizes the overall procedure of our proposed time-varying beamforming system, which consists of the time-frequency mask and attention weight estimation modules.

Fig. 2. Overall procedure of proposed time-varying beamforming system, which is constructed from the time-varying beamformer with the time-frequency mask and attention weight estimation modules.

2) Overview of Attention Module: Here, we briefly review the formulation of an attention module. Let \( q_{t} \in \mathbb{R}^{1 \times D^{K}} \), \( k_{t} \in \mathbb{R}^{1 \times D^{Q}} \), and \( v_{t} \in \mathbb{R}^{T \times D^{V}} \) be the vectors at time frame \( t \) called query, key, and value, respectively. Here, \( D^{KQ} \) denotes the dimension of the query and key, and \( D^{V} \) denotes the dimension of the value.

Given the sequence of the queries \( Q = \{ q_{t=1}, \ldots, q_{t=T} \} \in \mathbb{R}^{T \times D^{KQ}} \), keys \( K = \{ k_{t=1}, \ldots, k_{t=T} \} \in \mathbb{R}^{T \times D^{KQ}} \), and values \( V = \{ v_{t=1}, \ldots, v_{t=T} \} \in \mathbb{R}^{T \times D^{V}} \) as a matrix form, the output
of the attention module is computed as:
\[
A = \text{Weight}(Q, K) = \text{Softmax} \left( \frac{QK^T}{\sqrt{D}} \right),
\]
(14)
\[
Z = \text{Att}(A, V) = AV,
\]
(15)
where $\text{Weight}(Q, K)$ denotes the function computing the attention weights $A = \{a_{t=1}, \ldots, a_{t=T}\} \in \mathbb{R}^{T \times T}$, and $\text{Att}(A, V)$ denotes the function computing the attention output $Z = \{z_{t=1}, \ldots, z_{t=T}\} \in \mathbb{R}^{T \times D}$. Here, $a_t = \{a_{t,\nu=1}, \ldots, a_{t,\nu=\nu'}\} \in \mathbb{R}^{1 \times T}$ and $z_t \in \mathbb{R}^{1 \times D}$ denote the attention weight and output corresponding to query time $t$, respectively. $\text{Softmax}(\cdot)$ is the softmax function [43] that normalizes the attention weights over a key’s axis. A self-attention module is a special case of attention that uses the same features for the query, key, and values.

The attention module outputs the sum of the value features weighted by the attention weights as in (15). We can confirm that the computation of (15) is similar to that of (8) because we can reformulate (15) in a vector form as $z_t = \sum_{\nu=1}^{\nu'} a_{t,\nu} v_{\nu}$, where the value $v_{\nu}$ in (15) corresponds to the ISCM $\psi_{\nu,\nu'}$ in (8), the attention weight $a_{t,\nu}$ corresponds to the weight $c_{\nu,\nu'}$, and the output $z_t$ corresponds to the estimated SCM $\Phi_{\nu,\nu'}$.

Moreover, as seen from (14), the attention module determines the attention weights $a_{t,\nu}$ based on the dot-product similarity between queries and keys, and thus, the weight values become large when the input query and key features are similar. Therefore, the attention module would give larger weight values to the time frames where the positions of the target source speaker are similar, and it could thus automatically determine the frame regions suitable for computing the time-varying SCMs $\Phi_{\nu,\nu'}$ considering the position of the moving source speakers. Consequently, the self-attention-based NN can perform source tracking implicitly.

In more detail, to increase the representation capability, we adopted a stacked self-attention architecture [16], which consists of multiple self-attention modules as follows:
\[
Z_0 = \{\psi_{\nu,\nu'}^{1T}\}_{t=1},
\]
(16)
\[
A_i = \text{Weight}(K = Z_{i-1} W_i^K, Q = Z_{i-1} W_i^Q),
\]
(17)
\[
Z_i = \begin{cases} 
\text{Att}(A_i, V = Z_{i-1} W_i^V) & (i \neq I) \\
\text{Att}(A_I, V = Z_0) & (i = I)
\end{cases},
\]
(18)
where $Z_0$ is the input representation of the NN, and $Z_I$ is the output representation corresponding to the estimated time-varying SCMs $\Phi_{\nu,\nu'}^{\nu}$. $Z_i$ is the hidden representation at the $i$-th layer, and $I$ is the total number of layers. Here, $W_i^Q$, $W_i^K$, and $W_i^V$ are linear transformations associated with the query, key, and value, respectively. The learnable parameters of NN($\cdot$) are $\Lambda = \{W_i^Q, W_i^K, W_i^V\}_{i=1}^I$.

\section{Training Procedure}

We train the attention weight estimation NN using an objective function computed on the output of a mask-based beamformer so that it is possible to compute attention weights that are optimal for the beamforming of the moving source speaker; otherwise, it would be difficult to define the optimal target for the attention weights. We assume that a set of input and target signals $\{y, s\}$ is available for training the model, where $y \in \mathbb{R}^T$ is the $T$-length time-domain waveform of the observed noisy signal, and $s \in \mathbb{R}^T$ is its corresponding clean reverberant source signal. As the training objective, we adopted the scale-dependent signal-to-noise ratio (SNR) [44]. The SNR loss $\mathcal{L}$ is expressed as follows:
\[
\mathcal{L} = -10 \log_{10} \left( \frac{\|s\|^2}{\|s - \hat{s}\|^2} \right),
\]
(19)
where $\hat{s}$ denotes the time-domain waveform of the beamformed signal, which is computed based on the proposed scheme as described in Section IV-B. Here, we used a clean reverberant source (i.e., spatial image) as the network training target, because the MVDR beamformer in (3) is formalized to estimate the spatial image of a clean source [39] and we focused not on dereverberation but on noise reduction.

Through the training procedure, it is expected that the attention weight estimation networks learn to control the range for accumulating the ISCMs at each time step; consequently, the constructed beamformers can track the positions of the moving source speaker. We incorporate various moving source conditions in the training set to learn robust tracking capabilities. Such tracking behavior of the proposed scheme is visually analyzed in Section V-E2.

\subsection{Weight Smoothing}

Our preliminary experiments showed that while the proposed scheme is effective for improving the speech enhancement performance, e.g., SDR [45], it does not necessarily contribute to improving ASR performance. We hypothesized that this is probably due to the non-smoothness introduced by the frame-by-frame processing.

To mitigate this issue, we introduce a scheme to smooth the attention weights estimated with the NN as:
\[
\mathbf{c}_t^{\nu} = \frac{1}{\sum_{\tau=t-L'}^{t+L'} c_\tau^{\nu}},
\]
(20)
where $c_\tau^{\nu}$ is the smoothed version of the weight coefficients and $L'$ determines the number of frames used for the weight smoothing.

Here, (20) may look similar to the blockwise computation approach, since the summation of weights is performed over a window. However, the weights $c_\tau^{\nu}$ span the entire signal, unlike in blockwise processing, and it thus results in very different processing.

\section{Experiment}

\subsection{Experimental Conditions}

To evaluate the effectiveness of the proposed method, we created a new dataset of simulated moving sources in noisy conditions. The signals for the speech source were taken from the WSJ0 corpus [20] and those for the noise from the CHiME-3

\footnote{To simplify the description, we explain the single-head attention case, although we use a multi-head attention followed by a position-wise feed-forward network [16] for $i \neq I$ in our experiments.}
The CHiME-3 corpus contains noise signals recorded using a tablet device equipped with a rectangular microphone array with 6 channels, as illustrated in Fig. 3. From the 6-channel microphones, we excluded the second channel’s signal, which was captured by a microphone facing backward the tablet, and used the remaining five channels for the following multichannel experiments (i.e., $C = 5$).

We randomly selected the pair of speech and noise signals from the WSJ0 and CHiME-3 corpora, respectively, and mixed them at various SNR between 2 dB and 8 dB. We generated room impulse response (RIR) for moving sources using the gpuRIR simulation toolkit [21], which is based on the image method [46]. We used a randomly generated configuration (i.e., room geometry, array position, and source trajectory) for each simulated RIR. Fig. 4 shows an example of such a layout. In this experiment, we assumed that the room geometry was square and the source speaker was moving in a straight line in the room. As illustrated in Fig. 4, the start and end positions of the source trajectory are randomly sampled from the red area, and the array position is randomly sampled from the blue area. We set our simulation so that each moving speaker would start speaking an utterance at the start position and stop speaking at the end position. The speed of the moving source speaker is constant within an utterance, but varies across utterances. The minimum, maximum, and average source velocities [m/s] (moving distance [m]/utterance duration [s]) of the generated moving source signals are 0.01 m/s, 2.30 m/s, and 0.22 m/s, respectively. The reverberation time (T60) ranges from 0.1 to 0.3 s. Table I summarizes the configuration of the moving source simulation.

We created 30,000, 2,000, and 2,000 noisy speech signals for training, development, and evaluation sets, respectively. The speech sources for the training set were selected from WSJ0’s training set “si_tr_s.” Those for the development and evaluation sets were selected from WSJ0’s development set “si_dt_05” and evaluation set “si_et_05,” respectively. We generate noisy signals using the noise from the CHiME-3 corpus. We divided the noise sources in the CHiME-3 corpus into three subsets for training, development, and evaluation, containing 80%, 10%, and 10%, respectively, of the noise data of each environment (on a bus, in a cafe, pedestrian area, and street junction).

In addition to the above moving source dataset, we also created a non-moving source dataset as the additional evaluation set, which has exactly the same configuration as the moving source dataset (i.e., the pair of speech and noise sources and the RIR configurations) except that the source speaker position is fixed to the start position.

As the evaluation metrics, we used three speech enhancement measures; 1) the signal-to-distortion ratio (SDR) that permits time-invariant filters allowed distortions [45], 2) perceptual evaluation of speech quality (PESQ) [47], and 3) short-time objective intelligibility (STOI) [48]; in addition, we used one speech recognition measure, i.e., word error rate (WER). To compute the speech enhancement measures, we used the clean reverberant signals of the moving source speakers at the fifth channel as their references.

As the evaluation metrics, we used three speech enhancement measures; 1) the signal-to-distortion ratio (SDR) that permits time-invariant filters allowed distortions [45], 2) perceptual evaluation of speech quality (PESQ) [47], and 3) short-time objective intelligibility (STOI) [48]; in addition, we used one speech recognition measure, i.e., word error rate (WER). To compute the speech enhancement measures, we used the clean reverberant signals of the moving source speakers at the fifth channel as their references.

To evaluate the ASR performance, we created a deep neural network-hidden Markov model (DNN-HMM) hybrid ASR system [49] based on Kaldi’s CHiME-4 recipe [50]. The system was trained using the lattice-free maximum mutual information (MMI) criterion [51] with the noisy speech signals in the training set, and decoded with a trigram language model. The details of the system are shown in Kaldi’s recipe\(^2\).

\[\text{https://github.com/kaldi-asr/kaldi/tree/master/egs/chime4}\]
B. Experimental Configurations

For the time-varying attention weight estimation module described in Section IV, we adopted a self-attention-based network architecture that is similar to the one used by the Transformer encoder [16]. It consisted of stacked self-attention blocks, each of which was composed of the multi-head attention module followed by the position-wise feed-forward network. For the training loss of the attention weight estimation NN, we adopted the SNR loss shown in (19), where the enhanced signals are obtained by applying the beamforming filters to the observed signals in (2). The beamforming filter coefficients were obtained in (3) using the time-varying SCMs estimated by the proposed method. In the training stage, we used the “Wiener like” oracle time-frequency masks [52] to compute the ISCMs and optimized only the parameters of the attention weight estimation module (\(I^V\)) in Section IV-B2 based on the moving source dataset. In the testing stage, we used the estimated time-frequency masks, which is obtained by averaging the estimated time-frequency masks computed from each microphone signal separately [10]. We set the number of frames for weight smoothing \(L'\) in (20) to 7 for moving source dataset and to 9 for non-moving source dataset based on the WER scores on the development set.3

For the time-frequency mask estimation module, we adopted a CNN-based network architecture [53] that is similar to the time-domain audio separation network (TasNet) [54]. It accepts a single-channel signal and outputs the time-frequency masks for the speech source. The NN consists of stacked dilated convolution blocks. Unlike a previous related work [54], it operates in the STFT domain [53]. For the training loss of the mask estimation NN, we also adopted the SNR loss shown in (19), motivated by the report that the Conv-TasNet model with SNR loss achieved higher estimation performance [53], where the enhanced signals are obtained by applying the estimated time-frequency masks to the observed signals.

For the STFT computation, we used a Hanning window with a length and shift set at 64 ms and 16 ms, respectively. The configurations related to the STFT and network architecture are briefly summarized in Table II, where we follow the notations introduced in [16] for attention weight estimation network and [54] for mask estimation network, respectively. We tuned the learning rate for each module independently by monitoring the loss score of the development set after a few dozen training epochs.

| TABLE II |
|------------------|------------------|------------------|
| **Summary of Experimental Configurations** | **Configuration of attention weight estimation network** | **Configuration of mask estimation network** |
| | Number of attention heads \(g^{\text{head}}\) & 4 |
| | Dimension of attention layers \(d^{\text{dim-layer}}\) & 256 |
| | Dimension of feed-forward layers \(d^{\text{ff}}\) & 2048 |
| | Number of self-attention blocks \(N\) & 6 |
| | Batch size & 24 |
| | Learning rate & 5e-5 |
| | Optimization technique & Adam |
| | Total number of model parameters & 14.6 M³ |
| | Number of channels in bottleneck (B) & 256 |
| | Number of channels in conv blocks (H) & 512 |
| | Number of conv blocks in each repeat (X) & 8 |
| | Number of repeats (R) & 4 |
| | Batch size & 24 |
| | Learning rate & 1e-4 |
| | Optimization technique & Adam |
| | Total number of model parameters & 8.8 M |
| | Sampling frequency & 16 kHz |
| | Frame length & 64 ms |
| | Frame shift & 16 ms |
| | Window function & Hanning |

All of our investigations are based on MVDR beamformers, but the proposed approach could in principle also be used with other types of differentiable beamformers such as the multichannel Wiener filter (MWF) [39] or the generalized eigenvalue decomposition (GEVD) beamformer [55].

For all beamformers, we adopted the diagonal loading technique [56] to stabilize the matrix inversion of the noise SCM in (3), i.e., we replaced \(\Phi_{t,f}^N\) with \(\Phi_{t,f}^N + \epsilon \text{Tr}(\Phi_{t,f}^N) I\), where \(I \in \mathbb{R}^{C \times C}\) is an identity matrix and \(\epsilon\) is a very small regularization constant that controls the relative loading level. For all of the evaluated beamformers, we set the regularization constant \(\epsilon\) to 1e-10.

1) Time-Invariant, Online and Blockwise Baselines: We adopted tiv_mvdr, onl_mvdr, and blk_mvdr as baselines that compute the SCMs based on the heuristic attention weights described in Section III-C. We tuned the forgetting factor \(\alpha\) and block size parameter \(L\) for the online and blockwise MVDR implementations, by varying the values of these parameters in the ranges of \(\alpha = \{0.999, 0.99, 0.9, 0.7, 0.5\}\) and \(L = \{5, 10, 20, 30, 40, 50\}\), respectively. We set the forgetting factor \(\alpha\) in (6) to 0.999 and the block size parameter \(L\) in (7) to 50 for the moving source dataset, and we set the forgetting factor \(\alpha\) to 0.999 and the block size parameter \(L\) to 50 for the non-moving source dataset, as they achieved the best WER scores on the development set.

2) DOA-Based Baseline: We adopted the DOA (source localization)-based beamformer (doa_mvdr) [33] as the baseline for a conventional signal processing-based beamformer. Here, we estimate the steering vector \(h_{t,f} \in \mathbb{C}^C\) from the DOA information based on the plane wave assumption as [33]:

\[
 h_{t,f} = \begin{bmatrix} e^{i \frac{2 \pi}{L} \tau_1}, e^{i \frac{2 \pi}{L} \tau_2}, \ldots, e^{i \frac{2 \pi}{L} \tau_C} \end{bmatrix}^T, \quad (21)
\]

In this experiment, we used two separate networks for computing the speech and noise SCMs. Consequently, the total number of parameters is double, i.e., 29.2 M. However, we believe that we could greatly reduce the number of parameters by, e.g., sharing the same model for the speech and noise SCMs and/or using smaller networks.

C. Evaluated Beamformers

In this paper, we compare our proposed self-attention-based time-varying MVDR beamformer (att_mvdr) with time-invariant (tiv_mvdr), online (onl_mvdr), and blockwise (blk_mvdr) MVDR beamformers. In addition to the above beamformers, we also evaluated the MVDR beamformers based on the direction-of-arrival (DOA) (doa_mvdr) [33] and 2) neural network-based forgetting factor tuning (fgt_mvdr) [15].

3In this experiment, we used two separate networks for computing the speech and noise SCMs. Consequently, the total number of parameters is double, i.e., 29.2 M. However, we believe that we could greatly reduce the number of parameters by, e.g., sharing the same model for the speech and noise SCMs and/or using smaller networks.
where $\nu$ is the speed of sound, and $\tau_c$ is the propagation delay with respect to the origin for microphone $c$. $\tau_c$ is computed as $\tau_c = -\frac{1}{\nu} p_c^T \mathbf{u}$, where $p_c \in \mathbb{R}^3$ is a three-dimensional vector representing the location of microphone $c$ and $\mathbf{u} = [\cos \theta \cos \varphi, \sin \theta \cos \varphi, \sin \varphi]^T \in \mathbb{R}^3$ is a three-dimensional unit vector representing the direction of a target source $[57]$. Based on free-field steering vector $\mathbf{h}_{t,f}$, we computed speech SCM as $\mathbf{F}^S_{t,f} = \mathbf{h}_{t,f} \mathbf{h}_{t,f}^H$, and constructed an MVDR beamformer by substituting $\mathbf{F}^S_{t,f}$ for $\mathbf{F}^S_{t,f}$ in (3), which corresponds to the standard formalization of the MVDR beamformer given a (free-field) steering vector $[33, 39]$. In this experiment, instead of applying the source localization (DOA estimation) method, we computed the oracle DOA information for each time frame using the sequence of the source locations and the location of the microphone array provided by the RIR simulator. Because the oracle DOA is computed based on the locations of the source and microphone, it would correspond to the upper-bound for the DOA estimation.

Furthermore, for computing the noise SCM $\mathbf{F}^N_{t,f}$ in (3), we utilized a widely-used WebRTC voice activity detection (VAD) toolkit $[58]$ and computed the noise SCM based on the detected noise regions (i.e., regions filtered out as non-speech). In this experiment, we evaluated two types of noise SCMs computed based on the VAD results: 1) computed by applying VAD to the reference clean source and 2) computed by applying VAD to the noisy observation, which we refer to as doa_mvdr ($o_{\text{doa}} + \text{src\_vad}$) and doa_mvd ($o_{\text{doa}} + \text{obs\_vad}$), respectively. Because the former VAD (i.e., src_vad) is computed based on the reference clean source, it would also correspond to the oracle detection of the noise regions.

3) Neural Network-Based Baseline: We also compared our proposed method with the neural network-based forgetting factor tuning approach (fgt_mvdr) $[15]$. This approach is an alternative to estimate the weight coefficients (here, simply the forgetting factor) for the SCM computation with a neural network. It performs online computation of the SCMs in a recursive manner as:

$$\mathbf{F}^V_{t,f} = \alpha^V_{t,f} \mathbf{F}^V_{t-1,f} + \frac{1 - \alpha^V_{t,f}}{L_B} \sum_{t=bL_B}^{(b+1)L_B-1} \mathbf{Y}^T_{t,f} \mathbf{Y}_{t,f},$$

where $\alpha^V_{t,f} = \{\alpha^V_{bL_B+1,f}, \ldots, \alpha^V_{F,f}\}$ denotes the forgetting factor coefficients for the $b$-th block, and $L_B$ denotes the frame length of each block. We set the block size $L_B$ to 5 in the experiment by following $[15]$. In $[15]$, the forgetting factor $\alpha^V_{t,f}$ is estimated with a feed-forward neural network FF($\cdot$) that recursively inputs the beamformer’s output as,

$$\mathbf{F}^V_{t,f} = \text{FF}(\{\hat{\mathbf{S}}^\text{pre}_{t,f} \}_{bL_B+1,f}^{(b+1)L_B-1,F}),$$

where $\hat{\mathbf{S}}^\text{pre}_{t,f} = \mathbf{w}^\text{pre}_{t-1,f} \mathbf{Y}_{t,f}$ is the beamformed signal obtained with the beamforming filter from the previous time step, $\mathbf{w}^\text{pre}_{t-1,f}$.

The main differences between our proposed approach and $[15]$ are as follows. First, the formalization of $[15]$ cannot use the future frames to compute the SCMs for each time step and is thus restricted to online processing, while our approach can use all of the frames and can perform batch processing. Second, to compute the SCMs for the $t$-th frame, $[15]$ estimated the forgetting factor for each frequency (i.e., $\alpha^V_{t,f} = \{\alpha^V_{f=1,f}, \ldots, \alpha^V_{f=F}\}$), while our approach estimates the attention weight for each time frame (i.e., $\alpha^V_{t,f} = \{\alpha^V_{t,f=1}, \ldots, \alpha^V_{t,f=F}\}$).

We followed the implementation described in the previous work $[15]$ for the forgetting factor estimation module. It consists of a two-layer feed-forward network with 2048 hidden units using batch normalization and dropout techniques. It has 3.2 M parameters, which is significantly fewer than the 29.2 M parameters in our proposed attention weight estimation network. We also evaluated five-layer and eight-layer network configurations that have a number of model parameters closer to our proposed method (15.8 M and 28.4 M, respectively), but these architectures performed worse than the two-layer configuration. For network optimization, we adopted the Adam algorithm $[59]$ with an initial learning rate of 0.001. To focus on the difference in SCM computation, we adopted the MVDR formalization in (3) and the SNR loss in (19), and we also used the same estimated time-frequency masks used for the other evaluated beamformers.

D. Experimental Results for Moving and Non-Moving Source Datasets

Here, we compare our proposed self-attention-based time-varying MVDR beamformer (att_mvdr) with the five MVDR beamformers, i.e., doa_mvdr, tiv_mvdr, onl_mvdr, blk_mvdr, and fgt_mvdr on moving and non-moving source datasets. As a comparison, we also provide the results obtained by applying the time-frequency mask to the mixture without any beamforming (i.e., masking). In this experiment, all of the above mask-based beamformers (i.e., except doa_mvdr) are constructed with the same estimated time-frequency masks, which are estimated by the time-frequency mask estimation network in Section V-B.

Table III shows the speech enhancement (i.e., SDR, PESQ, STOI) and ASR (i.e., WER) performance measures for the non-moving and moving source datasets. First, in comparison with mask-based and DOA-based beamformers, we observe that the DOA-based beamformer (i.e., doa_mvdr) performs worse in terms of enhancement measures than mask-based beamformers even when using the oracle DOA and VAD information (i.e., o_{doa} + src_vad) for both non-moving and moving source conditions. This is probably because the plane wave assumption is not valid for the type of reverberant rooms used in our experiments.

Next, we focus on the results of the mask-based beamformers. Looking at the behavior in the non-moving source scenario (left side of Table III), we observe that masking and all conventional variants of MVDR improve the speech enhancement measures,
TABLE III
SDR [dB], PESQ, STOI (HIGHER IS BETTER), AND WER [%] (LOWER IS BETTER) FOR NON-MOVING AND MOVING SOURCE DATASETS

| Method               | non-moving source | moving source |
|----------------------|-------------------|---------------|
|                      | SDR ↑ PESQ ↑ STOI ↑ | WER ↓ | SDR ↑ PESQ ↑ STOI ↑ | WER ↓ |
| mixture              | 5.3 1.37 0.87     | 4.9 | 5.3 1.38 0.87     | 4.9 |
| doa_mvdr (o_doa + src_vad) | 9.9 1.99 0.91 | 3.2 | 7.7 1.97 0.91 | 3.2 |
| doa_mvdr (o_doa + obs_vad) | 7.8 1.77 0.88 | 6.3 | 6.0 1.80 0.89 | 3.8 |
| masking              | 14.8 2.40 0.95 | 5.7 | 14.7 2.40 0.95 | 5.8 |
| tiv_mvdr             | 15.1 2.31 0.96 | 2.9 | 11.4 2.14 0.93 | 3.8 |
| onl_mvdr             | 13.5 2.24 0.95 | 3.4 | 10.2 2.08 0.92 | 4.1 |
| blk_mvdr             | 13.0 2.19 0.95 | 3.1 | 11.4 2.11 0.93 | 3.8 |
| fgt_mvdr             | 12.5 2.17 0.94 | 4.0 | 12.4 2.18 0.94 | 4.0 |
| Proposed att_mvdr    | 17.8 2.73 0.97 | 3.4 | 16.7 2.69 0.96 | 3.8 |
| + weight_smooth      | 15.4 2.48 0.96 | 3.0 | 13.9 2.48 0.95 | 3.4 |

i.e., SDR, PESQ, and STOI. For ASR, masking degrades performance, probably because it induces distortions that are harmful to ASR [60]. All conventional beamformers improve ASR, and the best performance is obtained with tiv_mvdr. This result is reasonable because the RIRs are static for this dataset.

The proposed att_mvdr achieves higher SDR, PESQ, and STOI scores compared to tiv_mvdr and comparable WER score when applying the smoothing scheme of (20) (i.e., att_mvdr+smooth). This result suggests that even for non-moving situations, the proposed method can improve the computation of the SCMs, probably because it may better adapt to changing noise conditions [19].

We then look at the behavior of the proposed approach in the moving source scenario (right side of Table III). We observe that the performance of tiv_mvdr degrades significantly compared to the non-moving case, i.e., SDR degrades by 3.7 dB and there is a relative WER degradation of more than 20%. This result confirms the importance of considering source movements in the design of a beamformer. onl_mvdr and blk_mvdr achieve time-varying beamforming, but they do not contribute to improving speech enhancement and ASR scores compared to tiv_mvdr. This illustrates the difficulty of setting appropriate hyperparameters to effectively track the moving sources.

In contrast, the proposed att_mvdr successfully achieved higher SDR, PESQ, and STOI scores compared to tiv_mvdr. In addition, by applying the weight smoothing scheme of (20), the proposed system (i.e., att_mvdr+weight_smooth) also successfully improved the WER performance compared to the baseline systems. These results confirm that the proposed time-varying beamforming approach can mitigate the performance degradation caused by moving sources.

Finally, we compare the proposed att_mvdr with fgt_mvdr, which uses a neural network to estimate the forgetting factors. fgt_mvdr achieves improved performance compared to onl_mvdr (and tiv_mvdr) for the moving source dataset (about 1 dB SDR improvement and slight WER degradation), but the improvement is still smaller than that of proposed att_mvdr. These results show the effectiveness of our proposed formalization in computing the SCMs with time-varying attention weights as shown in (8).

E. Experimental Analyses for Behavior of Proposed Self-Attention-Based Time-Varying Beamformer

In the following experiments, we analyze the behavior of the proposed scheme. In these analyses, we used the oracle time-frequency masks to focus on the behavior of the attention weight estimation module.

1) Visualization of Attention Weights: We analyzed the behavior of our proposed self-attention-based time-varying attention weight estimation by visualizing the attention weights. Fig. 5 plots the attention weights of an utterance in the moving source dataset for (1) the speech and (2) the noise SCM estimations, respectively. If the value of the time-frequency masks $m_{\nu,t}^\nu$ is close to zero, the attention weights $c_{t,t'}$ can take arbitrary values.
shows the speech enhancement performance measured on the moving source dataset. The performance of a fully causal mask-based beamforming system. The number of trajectory points is set to 360. The SNR is set to 10 dB. The source velocity of the generated moving source signal is 2.19 m/s.

Fig. 7 shows an example of the beam patterns for a moving source, where beam patterns at eight time frames are shown. The black straight line denotes the actual direction of the source speaker at that frame, i.e., 30°, 60°, 70°, 90°, 120°, 130°, 210°, and 270°. The blue and red lines correspond to the beam patterns for 1 kHz and 2 kHz, respectively. Fig. 7 also plots the spectrograms of the enhanced, noisy observation, and clean reference signals to show the speech activity of the visualized utterance.

The first beam pattern on the left (i.e., 30°) corresponds to a region where the source is inactive. In this case, the beamformer does not show any clear directivity pattern. On the other hand, when the source speaker is active (e.g., 60° and 90°), we can confirm that the beamformer has a main lobe toward the direction of the source speaker. Moreover, we observe that the beam patterns change over time and follow the source positions. These visualizations suggest that the estimated beamforming filters of the proposed method can successfully track the positions of a moving source speaker.

F. Evaluation of Applicability for Sequential and Latency-Controlled Processing

In the above experiments, we focused on offline processing that utilizes all of the information within an utterance to compute the attention weight for each time frame \( t \), \( c_{t}^{\nu} \), as in (13). Here, to confirm the potential of the proposed framework for application to sequential (causal) or latency-controlled processing, we evaluate the proposed att_mvdr using various numbers of look-ahead frames for the self-attention computation. In this preliminary experiment, to allow comparison with the previous experiments, we used the same estimated time-frequency masks used for the evaluated beamformers in Table III, which is not a causal implementation. Consequently, the results may not fully reflect the performance of a fully causal mask-based beamforming system.

Table IV shows the speech enhancement performance measures (i.e., SDR, PESQ, STOI) for the moving source dataset. For each evaluated att_mvdr, the attention weight for the \( t \)-th frame is computed as

\[
c_{t}^{\nu} = \text{NN}^{\nu}(\{\psi^{\nu}_{\nu^t} \}_{t=1}^{L}, \Lambda^{\nu}),
\]

where \( c_{t}^{\nu} \) denotes the attention weight at time frame \( t \), and \( \text{NN}^{\nu}(\cdot) \) is a neural network trained to estimate the attention weight based on the estimated time-frequency mask, the feature activation, and the current time frame. The number of trajectory points is set to 360. The SNR is set to 10 dB. The source velocity of the generated moving source signal is 2.19 m/s.
where $L_A$ denotes the number of look-ahead frames for the self-attention computation. Note that $L_A = 0$ corresponds to causal processing and $L_A = \infty$ corresponds to batch processing. In this experiment, we trained the attention weight estimation network from scratch for each look-ahead setting.

From the table, we can confirm that even for the causal setting (i.e., $L_A = 0$), the proposed att_mvd achieved better speech enhancement performance compared to all of the baseline beamformers in Table III. Moreover, we observe that by increasing the number of look-ahead frames, the performance approaches that of batch processing (i.e., $L_A = \infty$). These results demonstrate the potential of the proposed framework not only for batch processing but also for sequential and latency-controlled processing.

VI. CONCLUSION

In this paper, we discussed the application of mask-based beamformers to moving source situations. We introduced a generalized view of conventional approaches for computing the SCMs of moving sources, which can be interpreted as a sum of ISCMs weighted by attention weights. We proposed using an NN to compute these attention weights and showed that the self-attention-based NN is a reasonable candidate for this task. We performed experiments showing the impact of moving sources on conventional beamformers. The results show that it was challenging to achieve high enhancement and ASR performance when a source was moving even with an online or blockwise implementation of the mask-based beamformer. In contrast, the proposed scheme uses an NN to predict optimal attention weights to compute the time-varying SCMs. This resulted in stable performance for both moving and non-moving conditions. These results demonstrate the potential of our proposed approach as well as the importance of addressing the moving source conditions. Future works should include application of this framework to more challenging conditions, such as dealing with moving interfering sources (i.e., multiple moving sources) and higher reverberation conditions, as well as extend the approach to low-latency processing by, for example, exploring a lightweight network architecture with lower memory and computation costs. Moreover, future works should include the investigation for other types of beamformer formalizations, such as (parameterized) MWF [39], GEVD [55], and the recent weighted power minimization distortionless response (WPD) convolutional beamformers [61].

REFERENCES

[1] B. D. Van Veen and K. M. Buckley, “Beamforming: A versatile approach to spatial filtering,” IEEE ASSP Mag., vol. 5, no. 2, pp. 4–24, Apr. 1988.
[2] M. Brandstein and D. Ward, Microphone Arrays: Signal Processing Techniques and Applications. Berlin, Germany: Springer, 2001.
[3] J. Benesty, J. Chen, and Y. Huang, Microphone Array Signal Processing, vol. 1. Berlin, Germany: Springer, 2008.
[4] S. Doclo, S. Gannot, M. Moonen, A. Spriet, S. Haykin, and K. R. Liu, “Acoustic beamforming for hearing aid applications,” in Handbook on Array Processing and Sensor Networks, Hoboken, NJ, USA: Wiley, 2010, pp. 269–302.
[5] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, “Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction,” Speech Commun., vol. 49, no. 7/8, pp. 636–656, 2007.
[6] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, “Far-field automatic speech recognition,” Proc. IEEE, vol. 109, no. 2, pp. 124–138, Feb. 2021.
[7] 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 Int. Conf. Acoust., Speech, Signal Process., 2018, pp. 6697–6701.
[8] J. Heymann, M. Bacchiani, and T. N. Sainath, “Performance of mask based statistical beamforming in a smart home scenario,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2018, pp. 6722–6726.
[9] J. Heymann, L. Drude, and R. Haeb-Umbach, “Neural network based spectral mask estimation for acoustic beamforming,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2016, pp. 196–200.
[10] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, “Improved MVDR beamforming using single-channel mask prediction networks,” in Proc. Ann. Conf. Int. Speech Commun. Assoc., 2016, pp. 1981–1985.
[11] 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 Int. Conf. Acoust., Speech, Signal Process., 2016, pp. 5210–5214.
[12] 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 Autom. Speech Recognit. Understanding, 2015, pp. 504–511.
[13] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, “The fifth ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2018, pp. 1561–1565.
[14] J. Malek, Z. Kodolský, and M. Boháč, “Block-online multi-channel speech enhancement using DNN-supported relative transfer function estimates,” IET Signal Process., vol. 14, pp. 124–133, 2020.

Fig. 7. Visualization of beam patterns for a source moving around a circle as shown in Fig. 6.
D.-C. Chang and B.-W. Zheng, “Adaptive generalized sidelobe canceler,” IEEE Trans. Signal Process., vol. 34, no. 3, pp. 276–280, Mar. 1986.

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 Int. Conf. Acoust., Speech, Signal Process., 2021, pp. 6089–6093.

Y. Xu, Z. Zhang, M. Yu, S.-X. Zhang, and D. Yu, “Generalized spatio-temporal RNN beamformer for target speech separation,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2021, pp. 3076–3080.

T. N. Sainath et al., “Reducing the computational complexity of multichannel microphone acoustic models with integrated feature extraction,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2016, pp. 1971–1975.

G. Li, S. Liang, S. Nie, W. Liu, and Z. Yang, “Deep neural network-based generalized sidelobe canceller for dual-channel far-field speech recognition,” Neural Netw., vol. 141, pp. 235–247, 2021.

D.-C. Chang and B.-W. Zheng, “Adaptive generalized sidelobe canceller beamforming with time–varying direction-of-arrival estimation for arrayed sensors,” IEEE Sensors J., vol. 20, no. 8, pp. 4403–4412, Apr. 2020.

R. Schmidt, “Multiple emitter location and signal parameter estimation,” IEEE Trans. Antennas Propag., vol. 30, no. 3, pp. 340–349, May 1980.

J. H. DiBlasé, H. F. Silverman, and M. S. Brandstein, “Robust localization in reverberant rooms,” in Microphone Arrays. Berlin, Germany: Springer, 2001, pp. 151–170.

S. Chakraborty and E. A. P. Habets, “Multi-speaker DOA estimation using deep convolutional networks trained with noise signals,” IEEE J. Sel. Topics Signal Process., vol. 13, no. 1, pp. 8–21, Mar. 2019.

C. Schymura et al., “PILOT: Introducing transformers for probabilistic sound event localization,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2021, pp. 2117–2121.

C. Evers et al., “The LOCATA challenge: Acoustic source localization and tracking,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 28, no. 6, pp. 1620–1643, 2020.

K. Kumatani, J. McDonough, and B. Raj, “Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors,” IEEE Signal Process. Mag., vol. 29, no. 6, pp. 127–140, Nov. 2012.

S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” IEEE Trans. Signal Process., vol. 45, no. 8, pp. 1614–1626, Aug. 2001.

T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, “Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 25, no. 4, pp. 780–793, Apr. 2017.

J. Jansky, Z. Koldovský, J. Malek, T. Kounovsky, and J. Cmjevac, “Auxiliary function-based algorithm for blind extraction of a moving speaker,” EURASIP J. Audio, Speech, Music Process., vol. 2022, no. 1, pp. 1–16, 2022.

N. Amor, J. Cmjevac, V. Kautsky, Z. Koldovský, and T. Kounovsky, “Blind extraction of moving sources via independent component and vector analysis: Examples,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2021, pp. 3725–3729.

Z. Koldovský, V. Kautsky, P. Tichavsky, J. Cmjevac, and J. Malek, “Dynamic independent component/vector analysis: Time-variant linear mixtures separable by time-invariant beamformers,” IEEE Trans. Signal Process., vol. 69, pp. 2158–2173, 2021.

M. Souden, J. Benesty, and S. Affes, “On optimal frequency-domain multichannel linear filtering for noise reduction,” IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 2, pp. 260–276, Feb. 2010.

D. Wang, “Time-frequency masking for speech separation and its potential for hearing aid design,” Trends Amplification, vol. 12, no. 4, pp. 332–353, 2008.

D. H. T. Vu and R. Haeb-Umbach, “Blind speech separation employing directional statistics in an expectation maximization framework,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2010, pp. 241–244.

M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, “A multichannel MMSE-based framework for speech source separation and noise reduction,” IEEE Trans. Audio, Speech, Lang. Process., vol. 21, no. 9, pp. 1913–1928, Sep. 2013.

I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR–half-baked or well done?,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2019, pp. 626–630.

E. Vincent, R. Gribonval, and C. Fovette, “Performance measurement in blind audio source separation,” IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” J. Acoust. Soc. Amer., vol. 65, no. 4, pp. 943–950, 1979.

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 Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2001, pp. 749–752.

C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

G. Hinton et al., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, Nov. 2012.

D. Povey et al., “The Kaldi speech recognition toolkit,” in Proc. IEEE Workshop Auton. Speech Recognit. Understanding, 2011.

D. Povey et al., “Purely sequence-trained neural networks for ASR based on lattice-free MMI,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2016, pp. 2751–2755.

H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, “Phase-sensitive or well done?,” in Proc. IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 49, no. 8, pp. 1614–1626, Aug. 2019.

Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

E. Warsitz and R. Haeb-Umbach, “Blind acoustic beamforming based on generalized eigenvalue decomposition,” IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 5, pp. 1529–1539, Jul. 2007.

W. Zhang, X. Chang, C. Boeddeker, T. Nakatani, S. Watanabe, and Y. Qian, “End-to-end dereverberation, beamforming, and speech recognition in a cocktail party,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 30, pp. 3173–3188, 2022.

S. Araki, H. Sawada, R. Mukai, and S. Makino, “DOA estimation for multiple sparse sources with arbitrarily arranged multiple sensors,” J. Signal Process. Syst., vol. 63, no. 3, pp. 265–275, 2011.

Accessed on: Oct. 1, 2022. [Online]. Available: https://github.com/wisemen/py-webrtcvad

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Representations, 2015.

S.-J. Chen, A. S. Subramanian, H. Xu, and S. Watanabe, “Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2018, pp. 1571–1575.
Tsubasa Ochiai (Member, IEEE) received the Ph.D. degree in information engineering from Doshisha University, Kyoto, Japan, in 2018. He is currently a Research Scientist with NTT Communication Science Laboratories, Kyoto. His research interests include speech and spoken language processing, such as automatic speech recognition, array signal processing, and machine learning. He is currently a Member of the Acoustic Society of Japan. He was the recipient of the Best Student Presentation Award from the Acoustical Society of Japan in 2014, Student Paper Award from the IEEE Kansai Section in 2016, Awaya Prize Young Researcher Award from the Acoustical Society of Japan in 2021, and Best Paper Runner-Up Award from the ACM Multimedia in 2022.

Marc Delcroix (Senior Member, IEEE) received the M.Eng. degree from the Free University of Brussels, Brussels, Belgium, and Ecole Centrale Paris, Paris, France, in 2003, and the Ph.D. degree from the Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan, in 2007. He is currently a Distinguished Researcher with NTT Communication Science Laboratories, NTT Corporation, Tokyo, Japan. He was a Research Associate with NTT Communication Science Laboratories from 2007 to 2008 and 2010 to 2012. Then he became a permanent Research Scientist with NTT Communication Science Laboratories in 2012. From 2015 to 2018, he was a Visiting Lecturer with the Faculty of Science and Engineering, Waseda University, Tokyo, Japan. His research interests include the various aspects of speech and audio processing, such as target speech and sound extraction, speech enhancement, robust speech recognition, model adaptation, and speaker diarization. He took an active part in the development of NTT robust speech recognition systems for the REVERB, CHiME 1, and CHiME 3 challenges, which all achieved the best performance results in the tasks. He was the recipient of the the 2006 Student Paper Award from the IEEE Kansai section, 2006 Sato Paper Award from ASJ, 2015 IEEE ASRU Best Paper Award Honorable Mention, 2016 ASJ Awaya Young Researcher Award, and ACM Multimedia 2022 best paper runner-up award. He is a Member of the IEEE Signal Processing Society Speech and Language Processing Committee, and an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. He was a Member of the organizing committee of the REVERB challenge 2014, ASRU 2017, and SLT 2022. He is a Member of ASJ.

Tomohiro Nakatani (Fellow, IEEE) received the B.E., M.E., and Ph.D. degrees from Kyoto University, Kyoto, Japan, in 1989, 1991, and 2002, respectively. He is a Senior Distinguished Researcher with NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan. Since joining NTT Corporation, as a Researcher in 1991, he has been investigating audio signal processing technologies for intelligent human-machine interfaces, including dereverberation, denoising, source separation, and robust ASR. He was a Visiting Scholar with the Georgia Institute of Technology, Atlanta, GA, USA, for a year in 2005, and Visiting Assistant Professor with the Department of Media Science, Nagoya University, Nagoya, Japan, from 2008 to 2017. From 2008 to 2010 he was an Associate Editor for the IEEE TRANSACTION ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. From 2009 to 2014, he was a Member of the IEEE Signal Processing Society Audio and Acoustics Technical Committee and from 2016 to 2021, Member of the IEEE SPS Speech and Language Processing Technical Committee. From 2011 to 2012, he was the Chair of the IEEE Kansai Section Technical Program Committee and from 2019 to 2020, Chair of the IEEE SPS Kansai Chapter. He was the Technical Program Co-Chair of the IEEE WASPAA-2007, Co-Chair of the 2014 REVERB Challenge Workshop, and General Co-Chair of the IEEE ASRU-2017. He is a Fellow of IEICE and Member of ASJ. He was the recipient of the 2008 IEICE Best Paper Award, 2009 ASJ Technical Development Award, 2012 Japan Audio Society Award, 2015 IEEE ASRU Best Paper Award Honorable Mention, 2017 Maejima Hisoka Award, and 2018 IWAENC Best Paper Award.

Shoko Araki (Fellow, IEEE) received the B.E. and M.E. degrees from The University of Tokyo, Tokyo, Japan, in 1998 and 2000, respectively, and the Ph.D degree from Hokkaido University, Sapporo, Japan, in 2007. She is currently a Senior Research Scientist with NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan, where she is currently leading the Signal Processing Research Group. Since she joined NTT in 2000, she has been engaged in research on acoustic signal processing, array signal processing, blind source separation, meeting diarization, and auditory scene analysis. She was formerly a Member of the IEEE SPS Audio and Acoustic Signal Processing Technical Committee during 2014–2019, the Vice Chair in 2022, and currently serves as its Chair. She has been a Board Member of Acoustical Society of Japan since 2017, and the Vice President of Acoustical Society of Japan since 2021. She also was a Member of the organizing committee of ICA 2003, IWAENC 2003, IEEE WASPAA 2007, HSCMA2017, IEEE WASPAA2017, IWAENC2018, IEEE WASPAA2021, and the evaluation Co-Chair of the Signal Separation Evaluation Campaign in 2008, 2010 and 2011. She was the recipient of the 19th Awaya Prize from Acoustical Society of Japan in 2001, Best Paper Award of the IWAENC in 2003, TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2004 and 2014, Academic Encouraging Prize from the Institute of Electronics, Information and Communication Engineers in 2006, Itakura Prize Innovative Young Researcher Award from ASJ in 2008, Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology, The Young Scientists’ Prize in 2014, IEEE SPS Best paper award in 2014, and IEEE ASRU 2015 Best Paper Award Honorable Mention in 2015. She has been an IEEE Fellow for contributions to blind source separation of noisy and reverberant speech signals since 2022.

[61] T. Nakatani and K. Kinoshita, “A unified convolutional beamformer for simultaneous denoising and dereverberation,” IEEE Signal Process. Lett., vol. 26, no. 6, pp. 903–907, Jun. 2019.