A Joint-Loss Approach for Speech Enhancement via Single-channel Neural Network and MVDR Beamformer

Zhi-Wei Tan, Anh H. T. Nguyen, Linh T. T. Tran, and Andy W. H. Khong
School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore
E-mail: zhiwei001@e.ntu.edu.sg, {nguyenhta, ttl.tran, andykhong}@ntu.edu.sg

***Abstract*—Recent developments of noise reduction involves the use of neural beamforming. While some success is achieved, these algorithms rely solely on the gain of the beamformer to enhance the noisy signals. We propose a framework that comprises two stages where the first-stage neural network aims to achieve a good estimate of the signal and noise to the second-stage beamformer. We also introduce an objective function that reduces the distortion of the speech component in each stage. This objective function improves the accuracy of the second-stage beamformer by enhancing the first-stage output, and in the second stage, enhances the training of the network by propagating the gradient through the beamforming operation. A parameter is introduced to control the trade-off between optimizing these two stages. Simulation results on the CHiME-3 dataset at low-SNR show that the proposed algorithm is able to exploit the enhancement gains from the neural network and the beamformer with improvement over other baseline algorithms in terms of speech distortion, quality and intelligibility.**

***Index Terms*—Neural beamforming, complex spectral mapping, speech enhancement, deep learning, joint-loss**

I. Introduction

Beamforming has played a key role in telecommunications [1], seismic application [2], speech localization [3] and speech enhancement [4–7]. Recent works on speech enhancement via beamforming have incorporated deep-learning techniques [8–18] and, which can broadly be classified as weight-prediction [11, 19], mask-based [8–13, 20], and spectral-mapping neural [14] beamforming techniques. Figures 1, and 2, illustrate the training and inference process of these approaches. In the weight-prediction approach, a neural network (NN) learns the mapping between the noisy signal and the beamforming weights. Mask-based and spectral-mapping approaches, on the other hand, predict the speech and noise components. Beamformer weights are then estimated using second-order statistics of the speech and noise components. Mask-based methods differ from spectral-mapping methods in that the former predict spectral masks while the latter directly predicts spectrograms of the target signal. To further improve the performance of neural beamforming, several methods enhances the training of the network by propagating the gradient through the beamforming operation [13, 20]. Such approaches have been shown to yield higher performance for speech recognition applications [13, 20]. A two-stage spectral-mapping neural beamforming has recently been proposed, where both noisy signal and beamformed multi-channel output of the first stage are used as inputs to the second stage [21]. These approaches, however, rely solely on its beamformer for speech enhancement.

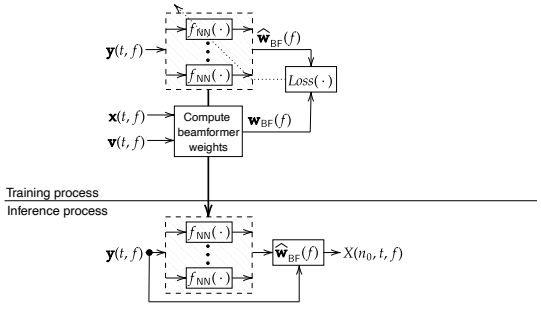
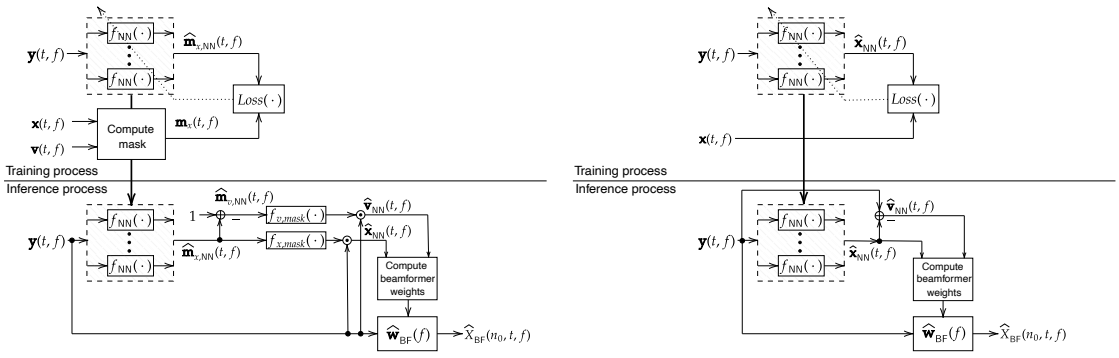


Fig. 1: Training and inference process of weight-prediction network approach.

We propose a new framework that incorporates two stages of speech enhancement. In the first stage, we employ SMoLnet [22] (a monoaural convolution neural network (CNN)) to estimate both the clean and noise signals. These signals are then used to estimate their corresponding covariance matrices for the second-stage minimum variance distortionless response (MVDR) beamformer [2, 6]. With such a cascaded structure, it is expected that any estimation errors present in the first stage will result in a sub-optimal performance for the second-stage beamformer. To address this issue, we propose a joint objective function that optimizes the outputs of *both* SMoLnet and MVDR. The weighting of this joint objective function is determined via a parameter that controls the trade-off between enhancing the multi-channel denoising capability of SMoLnet (which results in the improvement of performance for the second-stage beamformer), and optimizing the overall output (which inherently improves the training process of the neural-network by propagating the gradient through the beamforming

This work was supported within the STE-NTU Corporate Lab with funding support from ST Engineering and the National Research Foundation (NRF) Singapore under the Corp Lab@University Scheme (Ref. MRP14) at Nanyang Technological University, Singapore.



(a) Mask-based neural beamforming (b) Spectal-mapping neural beamforming

Fig. 2: Training and inference process of (a) mask-based and (b) spectral-mapping neural beamforming approach which optimizes based on neural network output.

operation). Simulations using the CHIME-3 dataset show that the proposed SMoLnet-MVDR algorithm achieves additional gain over both SMoLnet and the ablation version of SMoLnet-MVDR (SMoLnet-MVDR[a]), and the oracle MVDR beamformer, in terms of speech quality, distortion and intelligibility at low signal-to-noise ratios (SNRs).

II. Neural beamforming

A. Problem formulation

The multi-channel speech enhancement problem can be expressed in the time-frequency domain as

$$\begin{aligned} Y(n,t,f) &= G(n,t,f)\,S(t,f) + V(n,t,f) \\ &= X(n,t,f) + V(n,t,f)\,,\ n = 1,2,\ldots,N, \end{aligned} \quad (1)$$

where n is the microphone index and N is the number of microphones in the array, while t and f are the frame and frequency indices, respectively. In this signal model, the source signal $S(t,f)$ propagates through a channel that is sufficiently modeled by a convolutive filter[1] $G(n,t,f)$. The noise-free reverberant speech component $X(n,t,f)$ at the nth microphone is then contaminated by noise $V(n,t,f)$, resulting in a noisy speech signal $Y(n,t,f)$. We assume that the noise is uncorrelated with the source signal. The goal of speech enhancement algorithms is to achieve an accurate estimate of $X(n,t,f)$ by reducing noise component in $Y(n,t,f)$.

B. Mask-based and spectral-mapping neural beamforming

Mask-based neural beamforming methods achieve improvement over conventional parametric mask methods due to better estimation of the data-driven neural-network mask [23]. In these approaches, a denoising monoaural neural network estimates spectral masks via

$$\widehat{M}_{x,\mathsf{NN}}(n,t,f) = f_{\mathsf{NN}}(Y(n,t,f)), \quad (2)$$

where $f_{\mathsf{NN}}(\cdot)$ is the neural-network that maps each channel of the noisy signal to speech masks $\widehat{\mathbf{m}}_{x,\mathsf{NN}}(n,t,f) = \left[\widehat{M}_{x,\mathsf{NN}}(1,t,f),\ \ldots,\ \widehat{M}_{x,\mathsf{NN}}(N,t,f)\right]^{\mathsf{T}}$, and $(\cdot)^{\mathsf{T}}$ denotes the transpose operator. Noise masks can then be computed from the predicted speech mask [8] via

$$\begin{aligned} \widehat{\mathbf{m}}_{v,\mathsf{NN}}(t,f) &= 1 - \widehat{\mathbf{m}}_{x,\mathsf{NN}}(t,f) \\ &= \left[\widehat{M}_{v,\mathsf{NN}}(1,t,f),\ \ldots,\ \widehat{M}_{v,\mathsf{NN}}(N,t,f)\right]^{\mathsf{T}}. \end{aligned} \quad (3)$$

Alternatively, a neural-network can be trained such that it maps each channel of noisy signal to both speech and noise masks [23] by

$$[\widehat{M}_{x,\mathsf{NN}}(n,t,f), \widehat{M}_{v,\mathsf{NN}}(n,t,f)]^{\mathsf{T}} = f_{\mathsf{NN}}(Y(n,t,f)). \quad (4)$$

These masks are then used to compute a multi-channel NN-enhanced speech signal

$$\begin{aligned} \widehat{\mathbf{x}}_{\mathsf{NN}}(t,f) &= f_{x,\mathsf{mask}}\left(\widehat{\mathbf{m}}_{x,\mathsf{NN}}(t,f)\right) \odot \mathbf{y}(t,f) \\ &= \left[\widehat{X}_{\mathsf{NN}}(1,t,f),\ \ldots,\ \widehat{X}_{\mathsf{NN}}(N,t,f)\right]^{\mathsf{T}}, \end{aligned} \quad (5)$$

and an estimate of the input noise

$$\begin{aligned} \widehat{\mathbf{v}}_{\mathsf{NN}}(t,f) &= f_{v,\mathsf{mask}}\left(\widehat{\mathbf{m}}_{v,\mathsf{NN}}(t,f)\right) \odot \mathbf{y}(t,f) \\ &= \left[\widehat{V}_{\mathsf{NN}}(1,t,f),\ \ldots,\ \widehat{V}_{\mathsf{NN}}(N,t,f)\right]^{\mathsf{T}}, \end{aligned} \quad (6)$$

where $\mathbf{y}(t,f) = [Y(1,t,f),\ \ldots,\ Y(N,t,f)]^{\mathsf{T}}$, and $\odot$ is the Hadamard product operator. The post-masking schemes for the respective masks, $f_{x,\mathsf{mask}}(\cdot)$ and $f_{v,\mathsf{mask}}(\cdot)$ is optional, and is used to combine the multichannel masks to a single-channel mask. A non-exhaustive list of masking schemes includes median [9, 20], max [8] or mean [8, 11].

For spectral-mapping techniques, instead of predicting a mask, the neural network estimates the complex spectrogram of speech via [21, 22, 24, 25]

$$\begin{bmatrix} \widehat{X}_{\Re,\mathsf{NN}}(n,t,f) \\ \widehat{X}_{\Im,\mathsf{NN}}(n,t,f) \end{bmatrix} = f_{\mathsf{NN}}\left(\begin{bmatrix} Y_{\Re}(n,t,f) \\ Y_{\Im}(n,t,f) \end{bmatrix}\right), \quad (7)$$

[1]Here, we assume that the channel filter has a shorter length than the window length

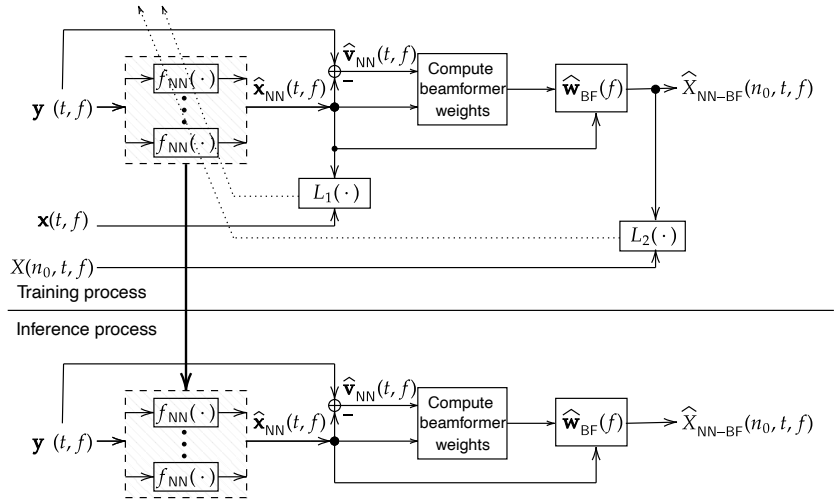


Fig. 3: Training and inference process of the proposed framework. In the training process, the neural network is optimized with a multi-channel objective L_1 and single-channel output objective L_2 as in (16). The output of the beamformer is decomposed to $X_{\mathsf{NN-BF}}(n_0)$.

where $f_{\mathsf{NN}}(\cdot)$ is a real-valued neural network that maps each channel of the noisy signal to speech component $\widehat{X}_{\mathsf{NN}}(n,t,f)$, and the subscript $\Re$ and $\Im$ indicate the real and imaginary components, respectively. Here, applying this monoaural neural network on each channel forms an output vector $\widehat{\mathbf{x}}_{\mathsf{NN}}(t,f) = \left[\widehat{X}_{\mathsf{NN}}(1,t,f),\ \ldots,\ \widehat{X}_{\mathsf{NN}}(N,t,f)\right]^{\mathsf{T}}$, where $\widehat{X}_{\mathsf{NN}}(n,t,f) = \widehat{X}_{\Re,\mathsf{NN}}(n,t,f) + j\widehat{X}_{\Im,\mathsf{NN}}(n,t,f)$.

Given estimates $\widehat{\mathbf{x}}_{\mathsf{NN}}(t,f)$ and $\widehat{\mathbf{v}}_{\mathsf{NN}}(t,f)$, the spatial covariance matrices of the NN-enhanced speech and estimated noise can be computed using

$$\widehat{\mathbf{\Phi}}_{xx,\mathsf{NN}}(f) = \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{x}}_{\mathsf{NN}}(t,f)\widehat{\mathbf{x}}_{\mathsf{NN}}^{\mathsf{H}}(t,f), \tag{8}$$

$$\widehat{\mathbf{\Phi}}_{vv,\mathsf{NN}}(f) = \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{v}}_{\mathsf{NN}}(t,f)\widehat{\mathbf{v}}_{\mathsf{NN}}^{\mathsf{H}}(t,f), \tag{9}$$

where T is the number of time frames and $(\cdot)^{\mathsf{H}}$ denotes the Hermitian transpose. These covariance matrices can then be employed to estimate the beamforming weights $\widehat{\mathbf{w}}_{\mathsf{BF}}(t,f)$, after which the beamformer output is obtained by

$$\widehat{X}_{\mathsf{BF}}(n_0,t,f) = \widehat{\mathbf{w}}_{\mathsf{BF}}^{\mathsf{H}}(f)\,\mathbf{y}(t,f), \tag{10}$$

where n_0 is the reference microphone index. In this framework, the neural-network parameters can be trained by propagating the gradient through the beamforming operation. As a result, a significant improvement on the final performance can be achieved [13, 20]. It is useful to note that the above algorithms rely solely on their beamformer, and their performance is upper-bounded by their oracle beamformers that are computed using the true covariance matrices. In the next section, we propose a framework to address this limitation.

III. Proposed SMoLnet-MVDR framework

The proposed two-stage speech enhancement framework is illustrated in Fig. 3. In the first stage, each channel of the noisy input signal is enhanced by a single-channel speech enhancement neural network, where the complex spectrogram of speech component is estimated via the spectral-mapping network in (7) and the complex spectrogram of input noise is defined as

$$\widehat{V}_{\mathsf{NN}}(n,t,f) = Y(n,t,f) - \widehat{X}_{\mathsf{NN}}(n,t,f). \tag{11}$$

In particular, we employ SMoLnet [22] for (7) because it is computationally efficient due to its significantly lower number of parameters compared to existing monoaural neural networks. In addition, its high-frequency resolution allows high fidelity in signal representation that, in turn, enhances the focusing ability of the beamformer. It is worth noting that other monaural speech enhancement methods can also be used.

In the second stage, an MVDR beamformer [2] is employed to further enhance the received signal at the output of the first stage. The MVDR beamformer weights are determined via the optimization criterion [6],

$$\begin{aligned}\min_{\widehat{\mathbf{w}}_{\mathsf{BF}}(f)}\ & \mathbb{E}_t\left\{\left|\widehat{V}_{\mathsf{NN-BF}}(n_0,t,f)\right|^2\right\}\\ \text{s.t}\ & \mathbb{E}_t\left\{\left|\widehat{X}_{\mathsf{NN-BF}}(n_0,t,f) - \widehat{X}_{\mathsf{NN}}(n_0,t,f)\right|^2\right\} = 0,\end{aligned} \tag{12}$$

where $\mathbb{E}_t\{\cdot\}$ denotes the sample mean along the time-frame axis, $\widehat{V}_{\mathsf{NN-BF}}(n_0,t,f) = \widehat{\mathbf{w}}_{\mathsf{BF}}^{\mathsf{H}}(f)\widehat{\mathbf{v}}_{\mathsf{NN}}(t,f)$, and $\widehat{X}_{\mathsf{NN-BF}}(n_0,t,f) = \widehat{\mathbf{w}}_{\mathsf{BF}}^{\mathsf{H}}(f)\widehat{\mathbf{x}}_{\mathsf{NN}}(t,f)$ denote the residual noise and enhanced signal at the output of the MVDR beamformer, respectively. Suppose that $\widehat{\mathbf{\Phi}}_{xx,\mathsf{NN}}(t,f)$ is rank-1, the

MVDR beamformer weights for (12) are given by

$$\mathbf{w}_{\mathsf{MVDR}}(f) = \frac{\widehat{\mathbf{\Phi}}^{-1}_{vv,\mathsf{NN}}(f)\,\widehat{\mathbf{\Phi}}_{xx,\mathsf{NN}}(f)}{\mathrm{tr}\Big(\widehat{\mathbf{\Phi}}^{-1}_{vv,\mathsf{NN}}(f)\,\widehat{\mathbf{\Phi}}_{xx,\mathsf{NN}}(f)\Big)}\mathbf{u}_{n_0}, \tag{13}$$

where $\mathrm{tr}(\cdot)$ is the trace operator, $\mathbf{u}_{n_0} = [0,\ \ldots\ 0, 1, 0,\ \ldots\ 0]^{\mathsf{T}}$ is a zero vector with a single value of one positioned at the n_0th element, and $(\cdot)^{-1}$ denotes the complex matrix inversion. Here, $\widehat{\mathbf{\Phi}}_{xx,\mathsf{NN}}(f)$ and $\widehat{\mathbf{\Phi}}_{vv,\mathsf{NN}}(f)$ are computed via (8) and (9), respectively. Since we implement all operations using real-valued operation, the complex-valued matrix inversion

$$\mathbf{A}^{-1} = (\mathbf{A}_{\Re} + \mathbf{A}_{\Im}\mathbf{A}_{\Re}^{-1}\mathbf{A}_{\Im})^{-1} - i(\mathbf{A}_{\Im} + \mathbf{A}_{\Re}\mathbf{A}_{\Im}^{-1}\mathbf{A}_{\Re})^{-1} \tag{14}$$

has been employed, where $\mathbf{A}_{\Re}$ and $\mathbf{A}_{\Im}$ are the real and imaginary components of a matrix $\mathbf{A}$, respectively [26]. Note that, in practice, $\mathbf{A}_{\Re}$ and $\mathbf{A}_{\Im}$ are diagonally loaded to avoid numerical issues.

Considering the output of the beamformer in (10), we can decompose it into the enhanced speech $\widehat{X}_{\mathsf{NN-BF}}(n_0,t,f)$, and residual noise $\widehat{V}_{\mathsf{NN-BF}}(n_0,t,f)$, as follows

$$\begin{aligned}\widehat{X}_{\mathsf{BF}}(n_0,t,f) &= \widehat{\mathbf{w}}^{\mathsf{H}}_{\mathsf{BF}}(f)\mathbf{y}(t,f)\\ &= \widehat{\mathbf{w}}^{\mathsf{H}}_{\mathsf{BF}}(f)\,\widehat{\mathbf{x}}_{\mathsf{NN}}(t,f) + \widehat{\mathbf{w}}^{\mathsf{H}}_{\mathsf{BF}}(f)\,\widehat{\mathbf{v}}_{\mathsf{NN}}(t,f)\\ &= \widehat{X}_{\mathsf{NN-BF}}(n_0,t,f) + \widehat{V}_{\mathsf{NN-BF}}(n_0,t,f).\end{aligned} \tag{15}$$

In constrast with existing neural beamforming technique, which utilizes $\widehat{X}_{\mathsf{BF}}(n_0,t,f)$ in (15) as the enhanced speech signal of the beamforming, the proposed SMoLnet-MVDR utilizes $\widehat{X}_{\mathsf{NN-BF}}(n_0,t,f)$ in (15) instead. Note that the extraction of $\widehat{X}_{\mathsf{NN-BF}}(n_0,t,f)$ is possible due to the proposed two-stage process, where $\widehat{\mathbf{x}}_{\mathsf{NN}}(t,f)$ is estimated in the first stage and $\widehat{\mathbf{w}}^{\mathsf{H}}_{\mathsf{BF}}(f)$ in the second stage.

With the above proposed framework, it is now possible to define a objective function that incorporates *both* multi-channel spectral mapping (Stage 1) and beamforming output (Stage 2) into account when optimizing for the neural network. More specficially, we define the proposed objective function as

$$L = \lambda L_1 + (1-\lambda)L_2,\ 0 \le \lambda \le 1 \tag{16}$$

where

$$L_1 = \mathbb{E}_{\mathrm{dataset}}\left\{\frac{1}{2}\left|\widehat{\mathbf{x}}_{\mathsf{NN}}(t,f) - \mathbf{x}(t,f)\right|^2\right\}, \tag{17}$$

$$L_2 = \mathbb{E}_{\mathrm{dataset}}\left\{\frac{1}{2}\left|\widehat{X}_{\mathsf{NN-BF}}(n_0,t,f) - X(n_0,t,f)\right|^2\right\}, \tag{18}$$

and $X(n_0,t,f)$ is the speech component in the reference channel. Here, $\mathbb{E}_{\mathrm{dataset}}(\cdot)$ denotes the sample mean over the whole training dataset. We, therefore, note that objective function L_1 optimizes the multi-channel denoising capability, which, in turn, enhances the estimation of spatial covariance matrices for a more accurate beamformer. It is important to note that, the beamforming weights for $\widehat{X}_{\mathsf{NN-BF}}(n_0,t,f)$ in (18) are computed for each signal during training and inference. Since $\widehat{X}_{\mathsf{NN-BF}}(n_0,t,f)$ contains distorted speech and noise due to imperfect spatial covariance, we introduce the objective function L_2 that serves as a feedback mechanism for the NN to learn its weights along with the beamforming operation such that $\widehat{X}_{\mathsf{NN-BF}}(n_0,t,f) \approx X(n_0,t,f)$. The variable λ therefore controls the amount of feedback to optimize the NN as shown in Fig. 3. With a large λ, the accuracy of beamformer is prioritized whereas with a small λ, the optimization of the overall output is prioritized.

The proposed framework is advantageous over the mask-based and spectral-mapping neural beamformer since it provides an additional gain from the NN to the final speech enhanced signal. In constrast to these methods which rely solely on beamforming, we employ the neural network to first reduce these noises before applying the beamformer. As such, the performance of the proposed framework is not upper-bounded by the performance of a beamformer. It is also useful to note that, this framework does not incur any additional computational cost compared to the neural beamforming framework.

IV. Simulation results

A. Simulation setup

Performance of the proposed framework is evaluated under low SNR conditions via the CHiME-3 dataset [27], and scaling it to low SNRs. We briefly describe this simulation setup that is designed for computer tablet usage in four noisy environments including pedestrian areas (PED), beside busy street intersection (STR), commercial cafe (CAF), and travelling on bus (BUS). For training, the source signals are the utterances in WSJ0 SI-84 [28] that are power normalized to the close-talking recordings in an acoustically isolated booth. For development and evaluation, the source signals are close-talking recordings in an acoustically isolated booth. To evaluate the algorithms at different SNRs, we utilized the simulated speech components which are formed by convolving each source signal with six time-varying filters. These six time-varying filters are estimated using time difference of arrival obtained from steered response power phase transform (SRP-PHAT) algorithm [29] of the six-channel microphones in the noisy environment [27]. They correspond to the acoustic impulse responses between the acoustic source and the microphones located on a tablet. The SNR is computed over all channels and the fifth channel is selected as the reference channel as it has the highest SNR. For the development set, environmental noise (PED, CAF, STR, and BUS) were added to simulated speech components to achieve SNR $= -15, -10,$ and -5 dB which corresponds to SNR $\approx -13.1, -8.1,$ and -3.1 dB on the fifth channel. For the evaluation set, SNR $= -15, -10, -5,\ 0,$ and 5 dB is obtained which corresponds to SNR $\approx -11.28, -6.28, -1.28,\ 3.7,$ and 8.7 dB on the fifth channel.

To train the network to be invariant to the loudness of noise without increasing the size of the training dataset, we scaled each training batch to an SNR drawn from a uniform distribution with $[-20,\ 0]$ dB which corresponds

TABLE I: SDR (dB) performance on the simulated CHiME-3 evaluation dataset for pedestrian (PED) and street (STR) noise types. Bolded values are best results excluding the oracle MVDR. The reference signal is recorded from the fifth channel (CH5). AVG denotes the average of each reported performance over SNR = $-15, -10, -5, 0$, and 5 dB. Results with asterisk (*) are better than the oracle MVDR beamformer.

Noise Type	PED						STR					
SNR (dB)	-15	-10	-5	0	5	AVG	-15	-10	-5	0	5	AVG
SNR (dB, CH5)	-10.99	-5.99	-0.99	4.01	9.01	-0.99	-11.10	-6.10	-1.10	3.9	8.9	-1.10
Unprocessed	-10.64	-5.86	-0.93	4.05	9.04	-0.87	-10.7	-5.94	-1.02	3.95	8.94	-0.96
Oracle MVDR [6]	-3.1	1.27	5.82	10.49	15.22	5.94	-1.5	2.93	7.49	12.07	16.53	7.5
NN-GEV [9]	-8.58	-3.44	0.76	3.65	5.01	-0.52	-8.05	-3.45	0.02	2.55	3.79	-1.03
SMoLnet (CH5)	-3.19	2.85*	7.09 *	10.64*	13.59	6.2*	-2.63	3.13*	7.49*	11.13	**14.07**	6.64
SMoLnet-MVDR[a]	-5.9	0.78	5.74	9.9	**12.61**	4.63	-4.47	1.79	6.83	10.83	12.88	5.57
SMoLnet-MVDR	**-0.28***	**4.96***	**8.68***	**11.37***	12.38	**7.42***	**-0.54***	**5.27***	**9.29***	**11.75**	12.27	**7.61***

TABLE II: SDR (dB) performance on the simulated CHiME-3 evaluation dataset for bus (BUS) and cafe (CAF) noise types. Bolded values are best results excluding the oracle MVDR. The reference signal is recorded from the fifth channel (CH5). AVG denotes the average of each reported performance over SNR = $-15, -10, -5, 0$, and 5 dB. Results with asterisk (*) are better than the oracle MVDR beamformer.

Noise Type	BUS						CAF					
SNR (dB)	-15	-10	-5	0	5	AVG	-15	-10	-5	0	5	AVG
SNR (dB, CH5)	-11.53	-6.53	-1.53	3.47	8.47	-1.53	-11.49	-6.49	-1.49	3.51	8.51	-1.49
Unprocessed	-11.16	-6.4	-1.47	3.5	8.5	-1.41	-11.09	-6.34	-1.42	3.55	8.54	-1.35
Oracle MVDR [6]	-2.62	1.86	6.49	11.14	15.74	6.52	-3.15	1.4	6.1	10.91	15.77	6.21
NN-GEV [9]	-7.56	-2.96	0.48	2.99	4.16	-0.58	-9.0	-3.98	0.3	3.3	4.72	-0.93
SMoLnet (CH5)	-1.45*	3.9*	7.98*	11.40*	14.24	7.21	-4.43	2.08*	6.57*	10.3	**13.31**	5.57
SMoLnet-MVDR[a]	-3.82	1.9*	6.54*	10.71	**14.26**	5.92	-6.21	0.62	5.7	9.9	12.52	4.51
SMoLnet-MVDR	**1.62***	**6.25***	**9.79***	**12.67***	14.16	**8.9***	**-1.63***	**4.09***	**7.86***	**10.61**	11.56	**6.5***

to approximately $[-18, 1.97]$ dB on the fifth channel. Each batch consists of eight shuffled training signals, each with a length of 0.64 s. Each signal was then transformed to the time-frequency domain using a short-time Fourier transform (STFT) with a 50% overlapping sinusoidal window, with each window being 2048 samples. We trained the proposed framework via the Adam optimizer [30] for five trials of randomly selected learning rates drawn from a log-uniform range between [5×10^{-5}, 0.001] and decreased its learning rate after ten epochs of development signal distortion ratio (SDR) [31] plateau. To verify the performance for a variety of λ, we varied λ randomly between [0.1, 1]. We exclude values lesser than 0.1 in the selection since a low value would result in an erroneous estimate of $\mathbf{x}(t, f)$, leading to poor signal and noise covariance matrices for the second-stage beamformer. We choose $\lambda = 0.30$ since it yielded the highest average SDR on the development set for evaluation.

We compared the proposed method with a neural masked-based network integrated with a generalized eigenvector beamformer. This NN-GEV [9] algorithm which utilizes a model with bi-directional long short-term memory (BLSTM) [32, 33] and a generalized eigenvector beamformer (GEV) with a distortion reduction filter [34]. In this method, we enhanced its reported training configuration for a fairer comparison as the dataset used a different SNR. The speech component of each training batch was then attenuated to achieve a randomly specified SNR. We then trained the model six times using the five previously-selected learning rates and the reported learning rate of 0.001. The model with best validation binary cross-entropy is chosen for testing. In addition, we utilized the maximum masking scheme [8] instead of the NN-GEV median masking scheme [9] as it empirically provides better validation in perceptual evaluation of speech quality (PESQ) [35, 36] and extended short-time intelligibility (ESTOI) [37]. We also set the thresholds for the noise mask at -30 dB along with the masks of unvoiced and voiced speech at -15 dB and -20 dB, respectively.

In addition to the above, we performed an ablation study for the proposed SMoLnet-MVDR framework. In this SMoLnet-MVDR[a] ablated version,

$$L^{\mathrm{a}} = \lambda L_1 + (1-\lambda)\mathbb{E}_{\mathrm{dataset}}\left\{\frac{1}{2}\left|\widehat{X}_{\mathsf{BF}}\left(n_0,t,f\right) - X\left(n_0,t,f\right)\right|^2\right\} \tag{19}$$

is the joint objective of the system with $\widehat{X}_{\mathsf{BF}}(n_0, t, f)$ being the estimated signal. As defined in (15), $\widehat{X}_{\mathsf{BF}}(n_0, t, f)$ consists of the desired speech component $\widehat{X}_{\mathsf{NN-BF}}(n_0, t, f)$, and the residual noise component $\widehat{V}_{\mathsf{NN-BF}}(n_0, t, f)$. For consistency, we performed the same training procedure as employed in the proposed method. To demonstrate the performance gain from the single-channel approach, we trained the SMoLnet on the single-channel objective based on signals from channel five and denoted it as SMoLnet (CH5). Furthermore, to show that the framework is not limited by the linear nature of the beamforming operation, we provide the oracle MVDR performance with its weights being computed for each signal as

$$\mathbf{w}_{\mathsf{MVDR}}(f) = \frac{\mathbf{\Phi}_{vv}^{-1}\left(f\right)\mathbf{\Phi}_{xx}\left(f\right)}{\mathrm{tr}\left(\mathbf{\Phi}_{vv}^{-1}\left(f\right)\mathbf{\Phi}_{xx}\left(f\right)\right)}\mathbf{u}_{n_0}, \tag{20}$$

where $\mathbf{\Phi}_{vv}(f) = \frac{1}{T}\sum_t^T \mathbf{v}(t,f)\mathbf{v}^{\mathsf{H}}(t,f)$ and $\mathbf{\Phi}_{xx}(f) = \frac{1}{T}\sum_t^T \mathbf{x}(t,f)\mathbf{x}^{\mathsf{H}}(t,f)$ denote the covariance matrices of the true speech signal $\mathbf{x}(t,f)$=$[X(1,t,f), \ldots X(\mathrm{N},t,f)]^{\mathsf{T}}$ and true noise signal $\mathbf{v}(t,f)$=$[V(1,t,f), \ldots V(\mathrm{N},t,f)]^{\mathsf{T}}$, respectively. The neural networks and the MVDR beamformer are

TABLE III: PESQ performance on the simulated CHiME-3 evaluation dataset for pedestrian (PED) and street (STR) noise types. Bolded values are best results excluding the oracle MVDR. The reference signal is recorded from the fifth channel (CH5). AVG denotes the average of each reported performance over SNR = $-15, -10, -5, 0$, and 5 dB. Results with asterisk (*) are better than the oracle MVDR beamformer.

Noise Type	PED						STR					
SNR (dB)	-15	-10	-5	0	5	AVG	-15	-10	-5	0	5	AVG
SNR (dB, CH5)	-10.99	-5.99	-0.99	4.01	9.01	-0.99	-11.10	-6.10	-1.10	3.9	8.9	-1.10
Unprocessed	1.14	1.25	1.49	1.81	2.15	1.57	1.1	1.29	1.62	1.97	2.33	1.66
Oracle MVDR [6]	1.46	1.59	1.83	2.14	2.49	1.9	1.47	1.66	1.94	2.28	2.63	2.0
NN-GEV [9]	1.21	1.33	1.71	2.08	2.4	1.75	1.09	1.42	1.82	2.2	2.51	1.81
SMoLnet (CH5)	1.15	1.59	2.09*	2.48*	2.75*	2.03*	1.26	1.72*	2.17*	2.52*	2.79*	2.09*
SMoLnet-MVDR[a]	1.31	1.61*	1.89*	2.18*	2.43	1.89	1.37	1.67*	1.97*	2.27	2.51	1.96
SMoLnet-MVDR	**1.39**	**1.91***	**2.36***	**2.66***	**2.83***	**2.23***	**1.46**	**1.94***	**2.35***	**2.64***	**2.81***	**2.24***

TABLE IV: PESQ performance on the simulated CHiME-3 evaluation dataset for bus (BUS) and cafe (CAF) noise types. Bolded values are best results excluding the oracle. The reference signal is recorded from the fifth channel (CH5). AVG denotes the average of each reported performance over SNR = $-15, -10, -5, 0$, and 5 dB. Results with asterisk (*) are better than the oracle MVDR beamformer.

Noise Type	BUS						CAF					
SNR (dB)	-15	-10	-5	0	5	AVG	-15	-10	-5	0	5	AVG
SNR (dB, CH5)	-11.53	-6.53	-1.53	3.47	8.47	-1.53	-11.49	-6.49	-1.49	3.51	8.51	-1.49
Unprocessed	1.12	1.37	1.71	2.08	2.44	1.74	1.17	1.26	1.48	1.79	2.13	1.57
Oracle MVDR [6]	1.6	1.79	2.05	2.36	2.7	2.1	1.47	1.61	1.86	2.19	2.54	1.93
NN-GEV [9]	1.18	1.56	1.98	2.37	2.64	1.95	1.17	1.26	1.62	2.03	2.4	1.7
SMoLnet (CH5)	1.28	1.76	2.23*	2.58*	2.85*	2.14*	1.14	1.58	2.07*	2.45*	2.72*	1.99*
SMoLnet-MVDR[a]	1.46	1.79	2.09*	2.39*	2.68	2.08	1.34	1.62*	1.91*	2.2*	2.45	1.9
SMoLnet-MVDR	**1.62***	**2.1***	**2.49***	**2.77***	**2.97***	**2.39***	**1.35**	**1.84***	**2.29***	**2.59***	**2.76***	**2.17***

constructed in PyTorch, a deep-learning framework where the gradient of each operation are computed using automatic differentiation [38].

B. Results

We evaluate the performance of the proposed SMoLnet-MVDR method in comparision with the oracle MVDR, NN-GEV, SMoLnet (CH5). We note that all speech enhancement methods mentioned in this paper provide performance improvement compared to the unprocessed signal. Their performance is compared in terms of SDR [31, 39], PESQ [35, 36] and ESTOI [37] for input SNR $= -15, -10, -5, 0$, and 5 dB in the presence of PED, STR, BUS and CAF noise. Tables I and II illustrate the performance of the proposed framework in terms of the SDR metric. It can be seen that at low SNRs of $-15, -10$, and -5 dB, the proposed framework achieves lower speech distortion than the all other baselines for all noise types. This result implies that the SDR performance of the proposed method at these SNRs is not upper-bounded by its oracle beamformer, where true speech and noise covariance matrices are being used. Furthermore, this result suggests that at these SNRs, the beamformer in the proposed method further reduces the speech distortion of the neural network in the first stage. This observation is expected since the SDR performance of the proposed SMoLnet-MVDR across all the considered noise types has outperformed SMoLnet-MVDR[a] (that relies solely on its beamforming enhancement), and similarly SMoLnet (CH5) (which relies solely on single-channel regression). At a higher SNR of 5 dB, the SDR performance of SMoLnet-MVDR is lower than that of oracle MVDR, SMoLnet-MVDR[a] and SMoLnet (CH5). This less-than-ideal performance may be attributed to the model mismatch between what are being used for training versus that used for testing for SMoLnet-MVDR. Nonetheless, the average SDR performance of the SMoLnet-MVDR has improved over SMoLnet-MVDR[a] and SMoLnet with a gain of 2.5 dB, and 1.4 dB, respectively. The lower distortion in the output also highlights the effectiveness of the proposed objective function at reducing distortion effectively for the two-stage enhancement.

Tables III and IV show the speech quality (PESQ) of the proposed method in comparison with other mentioned baselines. For speech quality, the proposed SMoLnet-MVDR outperforms all baseline methods at SNR $= -10, -5, 0$, and 5 dB. We observe that the proposed method outperforms the NN-GEV, SMoLnet (CH5) and SMoLnet-MVDR[a] for all SNRs in the evaluated range and for all noise types under consideration. It also outperforms the oracle MVDR for SNR $= -10, -5, 0$, and 5 dB and achieves a similar PESQ to the oracle MVDR for a very low SNR of -15 dB. In terms of average SNR, the proposed framework outperforms all the baselines and unprocessed signal for all noise types. In particular, it achieves an average improvement from 0.24 to 0.33 compared to the oracle MVDR and around 0.6 compared to the unprocessed signal. These observations confirm that the speech quality of the proposed framework is not upper-bounded by its oracle beamformer and that an additional gain can be achieved by using the proposed two-stage speech enhancement.

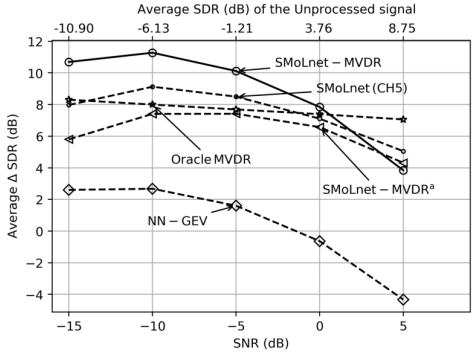
In addition, Tables V and VI show results in terms of speech intelligibility via the ESTOI measure for all methods. Overall, the oracle MVDR outperforms other methods for all noise types. However, in specific cases of PED and BUS noise with SNRs of -5, and 0 dB, the proposed SMoLnet-MVDR method outperforms other baseline methods including the oracle MVDR. For example, the SMoLnet-MVDR algorithm achieves a significant improvement of between 0.2 to 0.3

TABLE V: ESTOI performance on the simulated CHiME-3 evaluation dataset for pedestrian (PED) and street (STR) noise types. Bolded values are best results excluding the oracle MVDR. The reference signal is recorded from the fifth channel (CH5). AVG denotes the average of each reported performance over SNR = $-15, -10, -5, 0$, and 5 dB. Results with asterisk (*) are better than the oracle MVDR beamformer.

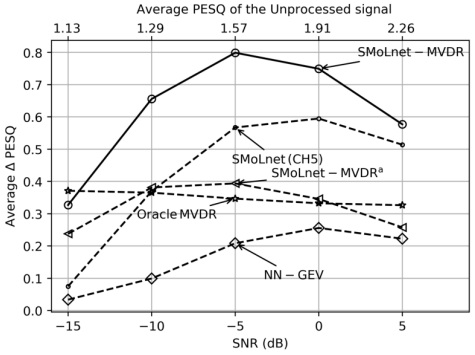
Noise Type	PED						STR					
SNR (dB)	-15	-10	-5	0	5	AVG	-15	-10	-5	0	5	AVG
SNR (dB, CH5)	-10.99	-5.99	-0.99	4.01	9.01	-0.99	-11.10	-6.10	-1.10	3.9	8.9	-1.10
Unprocessed	0.12	0.22	0.37	0.53	0.7	0.39	0.16	0.27	0.41	0.58	0.74	0.43
Oracle MVDR [6]	0.26	0.4	0.56	0.71	0.84	0.55	0.3	0.45	0.61	0.76	0.87	0.6
NN-GEV [9]	**0.18**	0.33	0.51	0.67	0.78	0.49	**0.22**	0.4	0.58	0.71	0.79	0.54
SMoLnet (CH5)	0.15	0.31	0.51	0.69	0.81	0.49	0.19	0.37	0.57	0.73	0.83	0.53
SMoLnet-MVDR[a]	0.16	0.34	0.54	0.7	0.82	0.51	0.19	0.37	0.57	0.73	**0.84**	0.54
SMoLnet-MVDR	0.17	**0.36**	**0.58***	**0.74***	**0.83**	**0.53**	0.21	**0.41**	**0.61**	**0.75**	0.83	**0.56**

TABLE VI: ESTOI performance on the simulated CHiME-3 evaluation dataset for bus (BUS) and cafe (CAF) noise types. Bolded values are best results excluding the oracle MVDR. The reference signal is recorded from the fifth channel (CH5). AVG denotes the average of each reported performance over SNR = $-15, -10, -5, 0$, and 5 dB. Results with asterisk (*) are better than the oracle MVDR beamformer.

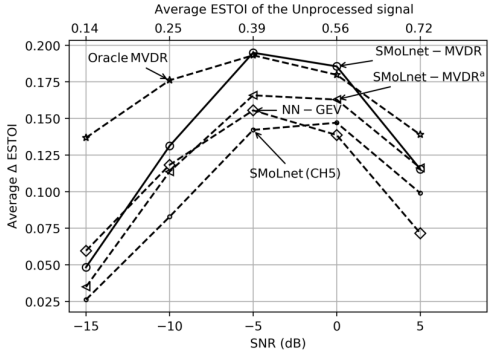
Noise Type	BUS						CAF					
SNR (dB)	-15	-10	-5	0	5	AVG	-15	-10	-5	0	5	AVG
SNR (dB, CH5)	-11.53	-6.53	-1.53	3.47	8.47	-1.53	-11.49	-6.49	-1.49	3.51	8.51	-1.49
Unprocessed	0.18	0.3	0.45	0.61	0.76	0.46	0.11	0.21	0.35	0.52	0.68	0.37
Oracle MVDR [6]	0.27	0.42	0.59	0.75	0.87	0.58	0.28	0.43	0.59	0.74	0.86	0.58
NN-GEV [9]	**0.26**	**0.45***	0.63	0.76	0.83	0.58	0.15	0.3	0.48	0.66	0.77	0.47
SMoLnet (CH5)	0.19	0.36	0.56	0.72	0.83	0.53	0.14	0.3	0.51	0.69	0.8	0.49
SMoLnet-MVDR[a]	0.21	0.39	0.58	0.74	0.86	0.56	0.15	0.35	0.55	0.72	0.83	0.52
SMoLnet-MVDR	0.24	0.43*	**0.62***	**0.77***	0.86	0.58	0.14	0.33	0.55	0.72	0.82	0.51



(a) Δ Average SDR (dB)



(b) Δ Average PESQ



(c) Δ Average ESTOI

Fig. 4: Improvement on average over all noise types (a) SDR (dB), (b) PESQ, and (c) ESTOI from unprocessed signal on the simulated CHiME-3 evaluation dataset. The average performance of the unprocessed signal over all noise types are placed above the each graph.

in terms of ESTOI compared to the oracle MVDR for the considered noise types and SNRs. The proposed method shows a modest reduction in ESTOI for BUS noise at very low SNRs of $-15, -10$ and -5 dB, and for PED, STR and CAF noise at SNR $= -15$ dB compared to the NN-GEV. However, the proposed method attains a much higher SDR and PESQ than NN-GEV. In addition, SMoLnet-MVDR achieves a modestly lower ESTOI for STR noise and PESQ compared to the SMoLnet-MVDR[a]. Nevertheless, the proposed SMoLnet-MVDR algorithm achieves significantly higher ESTOI than the NN-GEV, SMoLnet (CH5) and SMoLnet-MVDR[a] for PED and STR noise. A similar ESTOI is achieved compared to the oracle MVDR and the NN-GEV for BUS noise and only modestly lower ESTOI than SMoLnet-MVDR[a] for CAF noise.

Figure 4 summarizes the improvement in average performance over all noise types of the proposed method at SNR $= -15, -10, -5, 0,$ and 5 dB. On average, SMoLnet-MVDR outperforms other methods of SDR at SNR $= -15, -10, -5,$ and 0 dB and PESQ at SNR $= -10, -5, 0,$ and 5 dB. For speech intelligibility, it also achieves similar performance to that of the oracle MVDR at SNR $= -5$ and 0 dB and outperforms the non-oracle methods at SNR $= -10, -5,$ and 0 dB.

Figure 5 depicts the spectrograms of an evaluation utterance for female speaker on the street noise type for its unprocessed signal, and speech component on the fifth channel, and its enhanced signal by employing SMoLnet (CH5), SMoLnet-MVDR[a], and SMoLnet-MVDR. It is observed that the proposed SMoLnet-MVDR and SMoLnet (CH5) has greatly reduced the noise components in the noisy signal compared to SMoLnet-MVDR[a]. Specifically, the proposed SMoLnet-MVDR and SMoLnet (CH5) methods which involves the enhancement gain by the neural network achieve significantly lower amount of noise at low frequencies below 2 kHz than SMoLnet-MVDR[a] which involves the enhancement gain by the beamformer. This suggests that neural network approaches can be more effective at noise reduction compared to the beamforming approach. It is also useful to note that SMoLnet-MVDR was able to retain more high-frequency speech components above 4 kHz compared to SMoLnet (CH5) which can be observed at time 0.7, 1.8, 2.6, and 3.4 s. This suggests that the proposed SMoLnet-MVDR algorithm in conjuction with the joint objective function can utilize its beamformer to further to achieve lower speech distortion and higher speech intelligibility at high frequencies when compared to SMoLnet.

V. Conclusion

We propose a two-stage speech enhancement method that comprises a single-channel neural network and the MVDR beamformer. A joint objective function is formulated to minimize the speech distortion at each stage. Simulation results on the CHiME-3 dataset show that the proposed framework outperforms the NN-GEV, the SMoLnet (CH5) and the SMoLnet-MVDR[a] for all types of noise at SNRs from -15 to 0 dB. It also shows that the proposed framework outperforms its oracle beamformer at very low SNRs in terms of speech distortion and speech quality measures. This result implies that additional gain can be achieved by using the proposed two-stage framework in conjunction with the new joint objective function. In particular, the beamformer in the second stage can further enhance the output of the neural network in the first stage in terms of speech distortion, speech quality as well as intelligibility.

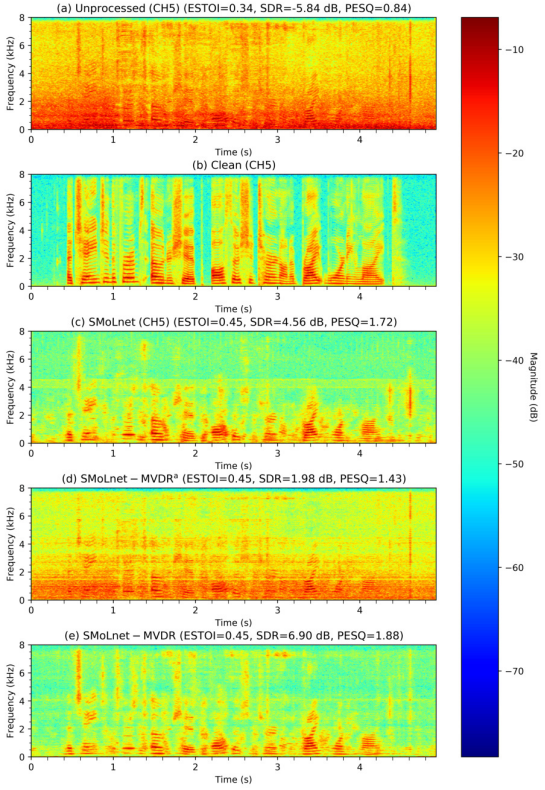


Fig. 5: Spectrograms of the evaluation utterance $f06\_444C0213\_STR$ for its (a) unprocessed signal and (b) speech component on the fifth channel, and after enhancement with the (c) SMoLnet (CH5), (d) SMoLnet-MVDR[a], and (e) SMoLnet-MVDR. The unprocessed signal has an SNR $= -10$ dB which corresponds to SNR $= -5.86$ dB at the fifth channel.

References

[1] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, 2016.

[2] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, pp. 1408–1418, 1969.

[3] V. V. Reddy, B. P. Ng, and A. W. H. Khong, "Insights into MUSIC-like algorithm," *IEEE Trans. Signal Process.*, vol. 61, pp. 2551–2556, 2013.

[4] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*, ser. Signals and Communication Technology. Berlin/Heidelberg: Springer, 2005.

[5] J. Benesty, J. Chen, Y. Huang, and J. Dmochowski, "On microphone-array beamforming from a MIMO acoustic signal processing perspective," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 1053–1065, 2007.

[6] 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, pp. 260–276, 2010.

[7] A. Ramírez López, N. Ono, U. Remes, K. Palomäki, and M. Kurimo, "Designing multichannel source separation based on single-channel source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 469–473.

[8] H. Erdogan, J. R. Hershey, S. Watanabe, M. Mandel, and J. L. Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. Interspeech*, 2016, pp. 1981–1985.

[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] 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.

[11] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE J. Sel. Topics Signal Process.*, vol. 11, pp. 1274–1288, 2017.

[12] A. S. Subramanian, C. Weng, M. Yu, S. Zhang, Y. Xu, S. Watanabe, and D. Yu, "Far-field location guided target speech extraction using end-to-end speech recognition objectives," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7299–7303.

[13] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 5325–5329.

[14] Z. Wang and D. Wang, "Multi-microphone complex spectral mapping for speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 486–490.

[15] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-attention dense u-net for multichannel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 836–840.

[16] S. Chakrabarty, D. Wang, and A. P. Habets, "Time-frequency masking based online speech enhancement with multi-channel data using convolutional neural networks," in *Proc. Int. Workshop Acoust. Signal Enhanc. (IWAENC)*, 2018, pp. 476–480.

[17] C. Liu, S. Fu, Y. Li, J. Huang, H. Wang, and Y. Tsao, "Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, pp. 1–1, 2020.

[18] S. Chakrabarty and E. A. P. Habets, "Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, pp. 787–799, 2019.

[19] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5745–5749.

[20] 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.

[21] Z. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1778–1787, 2020.

[22] Z.-W. Tan, A. H. T. Nguyen, and A. W. H. Khong, "An efficient dilated convolutional neural network for UAV noise reduction at low input SNR," in *Proc. Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2019, pp. 1885–1892.

[23] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *Proc. IEEE Work. Autom. Speech Recognit. Underst.*, 2015, pp. 444–451.

[24] K. Tan, J. Chen, and D. L. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, pp. 189–198, 2019.

[25] S.-W. Fu, T.-Y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2017, pp. 1–6.

[26] F. C. Chang. (2015) Complex matrix inversion by real matrix inversion. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/49373-complex-matrix-inversion-by-real-matrix-inversion

[27] Barker, Jon and Marxer, Ricard and Vincent, Emmanuel and Watanabe, Shinji, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop on Autom. Speech Recognit. Understanding*. IEEE, 2015, pp. 504–511.

[28] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete," *Linguist. Data Consortium, Philadelphia*, 2007.

[29] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays: Signal processing techniques and applications*, M. Brandstein and D. Ward, Eds. Springer-Verlag, 2001, ch. 8, pp. 157–180.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *in Proc. Int. Conf. Learn. Represent.*, pp. 127–142, 2015.

[31] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.

[33] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Sig. Proc.*, vol. 45, no. 11, p. 2673–2681, 1997.

[34] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 1529–1539, 2007.

[35] 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.

[36] "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU, ITU-T Rec. P. 862, 2000.

[37] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[38] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[39] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "mir_eval: A transparent implementation of common MIR metrics," in *Proc. Int. Soc. Music Inf. Retrieval*, 2014, pp. 367–372.