An MVDR-Embedded U-Net Beamformer for Effective and Robust Multichannel Speech Enhancement

Ching-Hua Lee[1], **Kashyap Patel**[2], **Chouchang Yang**[1], **Yilin Shen**[1], **Hongxia Jin**[1]

[1]Samsung Research America

[2]Department of ECE, University of Texas at Dallas

Contact Information:
Samsung Research America
665 Clyde Ave, Mountain View, CA 94043

Email: `chinghua.l@samsung.com`

Abstract

Objective:

- Develop a new deep neural network (DNN) model for achieving *effective* and *robust* multichannel speech enhancement (SE) simultaneously

Methods:

- Propose an **intra-MVDR embedded U-Net** to incorporate the merits of two popular DNN-based beamforming method types:
  - Type I: DNN direct beamformer (*effectiveness* in seen conditions)
  - Type II: Time-frequency (T-F) mask based statistical beamformer (*robustness* in unseen conditions)

Results:

- The proposed SE model demonstrates improved performance *which are not achievable by simply enlarging the baseline SE network of Type I or Type II*

1 Overview

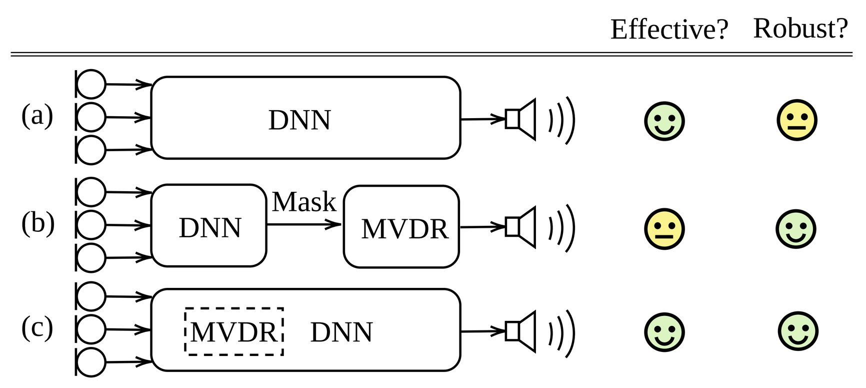

Figure 1: Illustration of different multichannel SE systems: (a) DNN direct beamformer; (b) DNN followed by statistical beamformer (e.g., MVDR); (c) MVDR-embedded DNN beamformer (proposed).

2 Background

2.1 Problem Formulation of Multichannel SE

- **Scenario:** one desired speech source and several interfering noise signals in a reverberant environment
- **Signal model:** T-F domain processing using the short-time Fourier transform (STFT) assuming an additive noise model:
  - N-mic array, the i-th microphone noisy signal STFT $\mathbf{X}_i \in \mathbb{C}^{F\times T}$ can be expressed as:

$$\mathbf{X}_i = \mathbf{S}_i + \mathbf{V}_i, \tag{1}$$

  $\forall i \in \{1, \ldots, N\}$, where $\mathbf{S}_i \in \mathbb{C}^{F\times T}$ and $\mathbf{V}_i \in \mathbb{C}^{F\times T}$ are the speech and noise components at microphone i, respectively.

- **Goal:** to recover the speech component $\mathbf{S} = \mathbf{S}_r$ of a reference microphone $r \in \{1, \ldots, N\}$ given the noisy $\mathbf{X}_1, \ldots, \mathbf{X}_N$

2.2 Type I: DNN direct beamformers (direct BF)

- The DNN $f_\theta(\cdot)$ is utilized to *imitate the beamforming processes for directly predicting the clean speech* $\mathbf{S}$, trained by minimizing some clean signal reconstruction loss:

$$\min_\theta \quad \mathcal{L}\left(\mathbf{S}, \hat{\mathbf{S}} = f_\theta(\mathbf{X}_1, \ldots, \mathbf{X}_N)\right) \tag{2}$$

- Effective as the model learns the direct noisy-clean mapping from data
- May not generalize adequately to unseen noise types and acoustic conditions not presented in training data

2.3 Type II: T-F Mask-Based Neural Beamformer

- The DNN $f_\theta(\cdot)$ is used to predict T-F mask $\mathbf{M}_s, \mathbf{M}_v \in \mathbb{C}^{F\times T}$ that represent the speech and noise T-F pattern:

$$\min_\theta \quad \mathcal{L}\left(\mathbf{M}_\gamma, \hat{\mathbf{M}}_\gamma = f_\theta(\mathbf{X}_1, \ldots, \mathbf{X}_N)\right), \quad \gamma = \{s, v\}, \tag{3}$$

  which are subsequently used to assist conventional beamformers, e.g., MVDR, based on estimating signal & noise statistics $\hat{\mathbf{S}} = g_{\text{mvdr}}([\mathbf{X}_1, \ldots, \mathbf{X}_N], [\mathbf{M}_s, \mathbf{M}_v])$

- Generalize better to unseen acoustic and noise conditions as the DNN only has to estimate the intermediate masks
- However, the overall SE performance is often bounded by the later statistical component (MVDR)

3 Proposed Model

3.1 Intra-MVDR module within direct BF network

The proposed model features intra-MVDR modules embedded in the U-Net direct BF (Figure 2). Here, MVDR is integrated as a network module and all the learnable parameters are jointly optimized for clean signal reconstruction

3.2 Exploiting MVDR-filtered signals at all mics

Each intra-MVDR module consists of:

[T-F mask estimation network → mask-based MVDR]

as Figure 3 illustrates, and performs MVDR for ALL mics

3.3 Multi-scale beamforming with intra-MVDR

Intra-MVDR naturally fits into a multi-scale design within the U-Net to better exploit coarse- and fine-grained spatial features from various resolutions

3.4 Combine MVDR-filtered signals at final output

The MVDR-filtered signals $\mathbf{Z}_i$ are included in the final filtering stage at the model output to help improve signal reconstruction

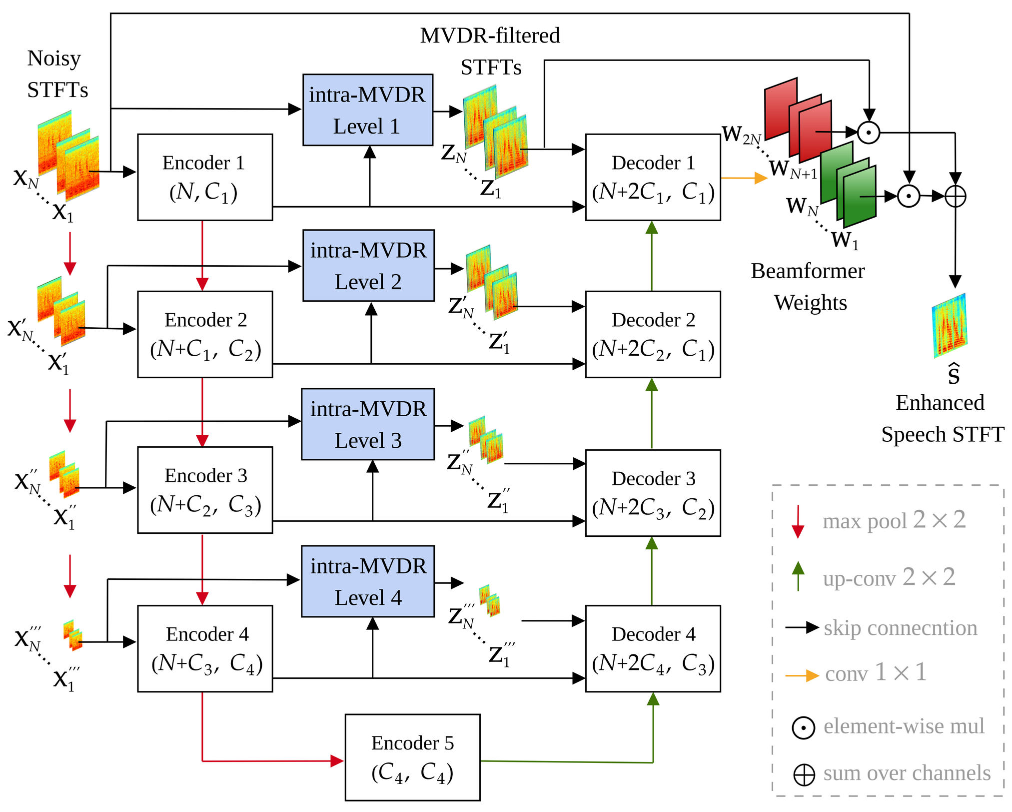

Figure 2: The proposed MVDR-embedded U-Net beamformer for SE.

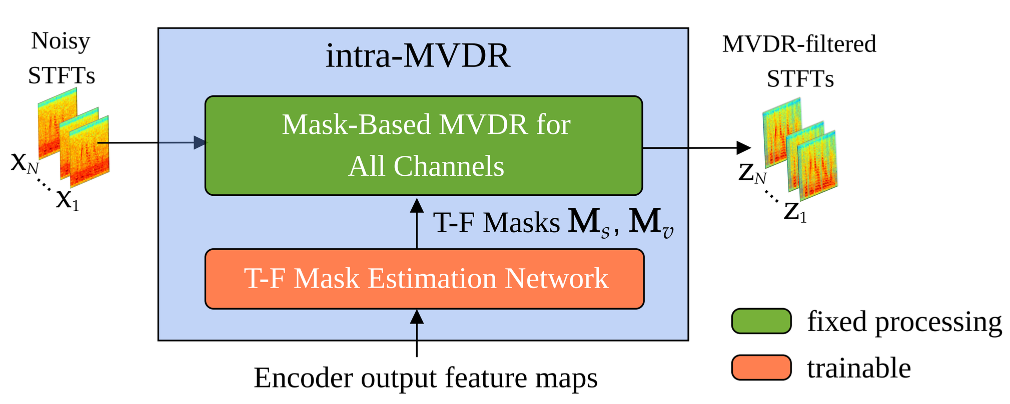

Figure 3: The proposed intra-MVDR module (Level 1) in details.

4 Experiments

We compare the SE performance of the following 3 cases based on using the same backbone (1.27M) U-Net model:

i) **Direct BF (Figure 1 (a)):** the U-Net is trained to directly estimate the clean speech

ii) **Mask-based MVDR (Figure 1 (b)):** the U-Net is trained to estimate the speech and noise ideal ratio masks

iii) **Direct BF w/ intra-MVDR (Figure 1 (c)):** the proposed intra-MVDR module(s) embedded in the U-Net direct BF

Datasets:

- CHiME-3 (for results in Table 1, Figure 4, Table 2)
- AVSpeech + Pyroomacoustics (for results in Table 3)

Results:

Table 1: Comparison of different multichannel SE schemes. For the direct BF and mask-based MVDR approaches we also show results for a larger (1.62M) U-Net models. For our method we present results for incorporating the intra-MVDR modules at different levels into the base (1.27M) U-Net model.

Methods		# Params	PESQ	STOI	SNR
Direct BF	(base)	1.27M	2.39	0.962	17.76
	(larger)	1.62M	2.44	0.965	18.31
Mask-based MVDR	(base)	1.27M	2.00	0.966	16.67
	(larger)	1.62M	2.01	0.966	16.81
Oracle MVDR		-	2.01	0.970	18.42
Direct BF w/ intra-MVDR	Level 1	1.30M	2.55	0.970	18.93
	Levels 1,2	1.38M	2.57	0.973	20.43
	Levels 1,2,3	1.47M	2.60	**0.974**	**20.80**
	Levels 1,2,3,4	1.56M	**2.64**	**0.974**	20.63

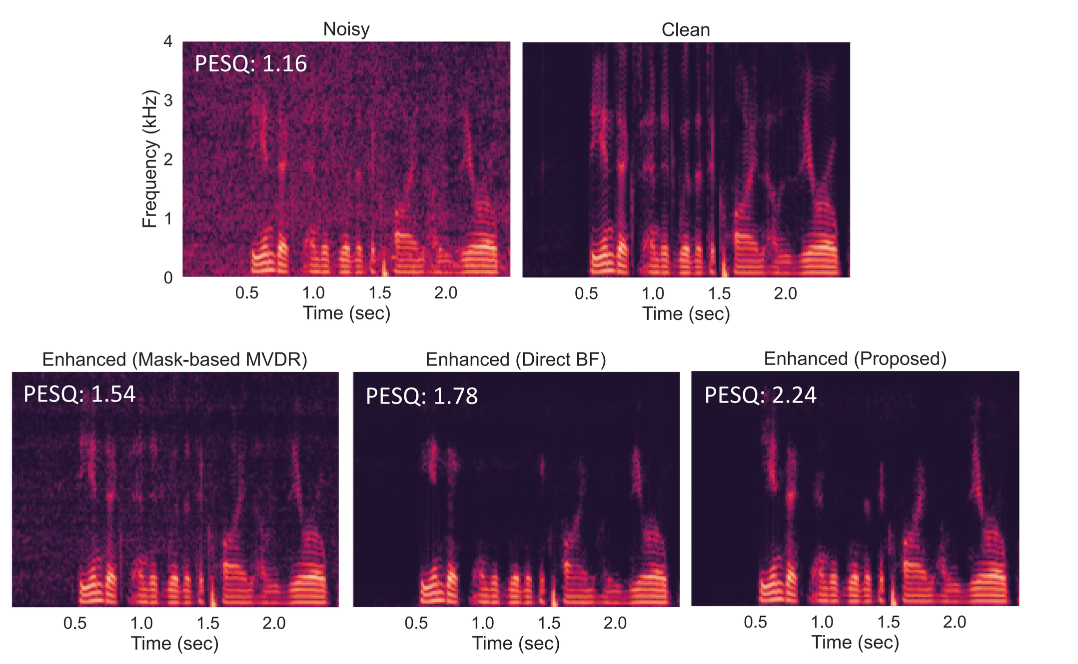

Figure 4: Visualization of SE outputs. The proposed method has less residual noise while preserving more speech components, achieving the best quality.

Table 2: Comparison with existing SE model for ASR.

Front-ends	# Params	WER / CER (%) ASR Model 1	ASR Model 2	ASR Model 3
Unprocessed	-	7.40 / 4.25	9.18 / 5.64	16.75 / 8.28
FaSNet	2.76M	5.21 / 2.63	5.65 / 3.41	10.20 / 4.87
Proposed	1.56M	**3.81 / 1.96**	**3.54 / 2.33**	**6.31 / 3.06**

Table 3: PESQ scores for comparing effectiveness on test data with seen room/noise conditions and robustness to unseen conditions.

Methods	# Params	Seen Cond.	Unseen Cond.
Noisy	-	1.21	1.22
Mask-based MVDR	1.62M	1.71	1.55
Direct BF	1.62M	2.02	1.66
Proposed	1.56M	**2.13**	**1.76**

5 Conclusion

We presented a novel integration of DNN direct beamforming and mask-based statistical beamforming by introducing the intra-MVDR module embedded in a U-Net design. The new model encompasses the merits of the two method types, efficiently improving SE effectiveness and robustness to various conditions.