Improved Mask-Based Neural Beamforming for Multichannel Speech Enhancement by Snapshot Matching Masking

Ching-Hua Lee, Chouchang Yang, Yilin Shen, Hongxia Jin

Samsung Research America (SRA)

The 48th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)
@ Rhodes Island, Greece
June 2023

SAMSUNG RESEARCH AMERICA

Introduction

- Time-frequency (T-F) mask-based neural beamformers have been widely adopted for multichannel speech enhancement (SE) tasks, which leverage deep neural networks (DNNs) to predict T-F masks for estimating the speech and noise *power spectral density (PSD)* matrices required in statistical beamforming algorithms.

- However, existing approaches train the DNNs to estimate some pre-defined masks, e.g., the ideal binary mask (IBM) and ideal ratio mask (IRM), that are not based on multichannel characteristics but on single-channel SE solutions, thus *lacking direct connection to the PSD estimation.*

- In this work, we propose a new masking strategy to predict the **Snapshot Matching Mask (SMM)** that aims to minimize the distance between the predicted and the true signal snapshots, thereby *estimating the PSD matrices in a more systematic way*, to achieve improved mask-based neural beamforming.

Problem Formulation

- **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:

Let f, t stand for the frequency and time frame indexes (total: F bins and T frames), the i-th microphone noisy signal STFT $\mathbf{X}_i \in \mathbb{C}^{F \times T}$ of an N-microphone array:

$$X_i(f,t) = S_i(f,t) + V_i(f,t), \tag{1}$$

$\forall f, t$, where $S_i(f,t)$ and $V_i(f,t)$ are the speech and noise components received by microphone i, respectively.

- **Goal:** to recover the speech component $\mathbf{S}_r \in \mathbb{C}^{F \times T}$ of a reference microphone $r \in \{1, \ldots, N\}$ given the N noisy signals $\mathbf{X}_1, \ldots, \mathbf{X}_N$

T-F Mask-Based Neural Beamformer

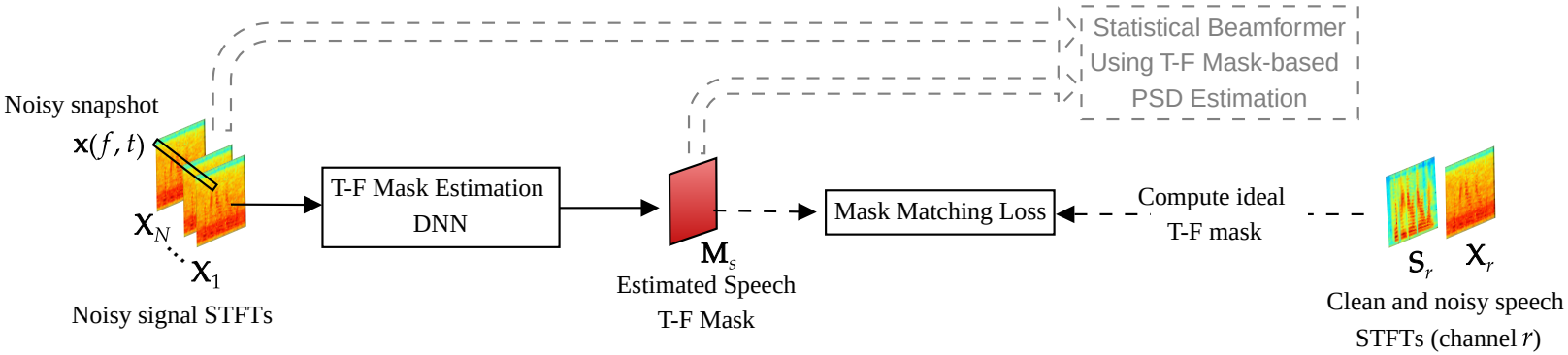

Figure: T-F mask-based neural beamformer: T-F mask estimation DNN (learning-based) followed by statistical beamformer (model-based).

- The T-F mask estimation DNN is utilized to predict some pre-defined ideal T-F masks that are subsequently leveraged to obtain an estimate of $\mathbf{\Phi}_s(f,t) = \mathrm{E}[\mathbf{s}(f,t)\mathbf{s}^H(f,t)]$ and $\mathbf{\Phi}_v(f,t) = \mathrm{E}[\mathbf{v}(f,t)\mathbf{v}^H(f,t)]$, the speech and noise PSD matrices, where $\mathbf{s}(f,t) = [S_1(f,t),\ldots,S_N(f,t)]^T$ and $\mathbf{v}(f,t) = [V_1(f,t),\ldots,V_N(f,t)]^T$ are the speech and noise snapshots.
- The statistical beamformer first computes the beamformer filter weights $\mathbf{w}(f,t) = g(\mathbf{\Phi}_s(f,t), \mathbf{\Phi}_v(f,t))$, typically as a function g of the speech and noise PSD matrices which are required to be estimated, and then applies the filter to the noisy snapshot as $\hat{S}_r(f,t) = \mathbf{w}^H(f,t)\mathbf{x}(f,t)$ to denoise.

Issues with Existing Masks

- E.g., the speech PSD can be estimated by recursive averaging:

$$\mathbf{\Phi}_s(f,t) = \lambda_s \mathbf{\Phi}_s(f,t-1) + M_s(f,t)\mathbf{x}(f,t)\mathbf{x}^H(f,t), \tag{2}$$

where $\lambda_s \in (0,1]$ is the forgetting factor and $M_s(f,t)$ is the DNN output mask to predict some ideal masks, e.g., the IBM and IRM:

$$M_s^{\text{IBM}}(f,t) = \begin{cases} 1, & \text{if } \frac{|S_r(f,t)|}{|V_r(f,t)|} > C \\ 0, & \text{otherwise} \end{cases}, \quad M_s^{\text{IRM}}(f,t) = \frac{|S_r(f,t)|}{|X_r(f,t)|}. \tag{3}$$

- However, the derivation of these T-F masks is not based on multichannel characteristics but on single-channel SE solutions.
- As a result, they lack direct relation to the PSD matrix estimation task that should account for all microphone channels jointly, and hence there is still room for improvement.

Proposed Method

Snapshot Matching Mask (SMM) estimation framework

$$M_s(f,t) = \underset{M\in\mathbb{C},|M|\leq 1}{\arg\min}\ \mathcal{L}(M\mathbf{x}(f,t), \mathbf{s}(f,t)), \tag{4}$$

$\forall f,t$, where $\mathcal{L}(\cdot,\cdot)$ is some measure of the difference between the estimated snapshot $M\mathbf{x}(f,t)$ and the clean speech snapshot $\mathbf{s}(f,t)$.

- We can see that the T-F mask $\mathbf{M}_s \in \mathbb{C}^{F\times T}$ given by (4) leads to an estimate of the speech signal snapshot $\mathbf{s}(f,t)$, i.e.,

$$\hat{\mathbf{s}}(f,t) \triangleq M_s(f,t)\mathbf{x}(f,t) \approx \mathbf{s}(f,t), \tag{5}$$

  $\forall f,t$. By matching the snapshots, we can better estimate $\mathbf{\Phi}_s(f,t) = \mathrm{E}[\mathbf{s}(f,t)\mathbf{s}^H(f,t)]$ by leveraging

$$\hat{\mathbf{s}}(f,t)\hat{\mathbf{s}}^H(f,t) \approx \mathbf{s}(f,t)\mathbf{s}^H(f,t). \tag{6}$$

- We also bound the magnitude of the mask within the unit circle for avoiding the difficulty of optimizing from an infinite search space.

SMM vs. Existing Masks

Proposed SMM: exploiting inter-channel correlation of multichannel target clean speech signals for network learning via a complex-valued masking scheme to directly minimize the distance between the estimated and the true signal snapshots

Existing T-F masking: only using the single reference channel information as the (real-valued) target mask for network learning

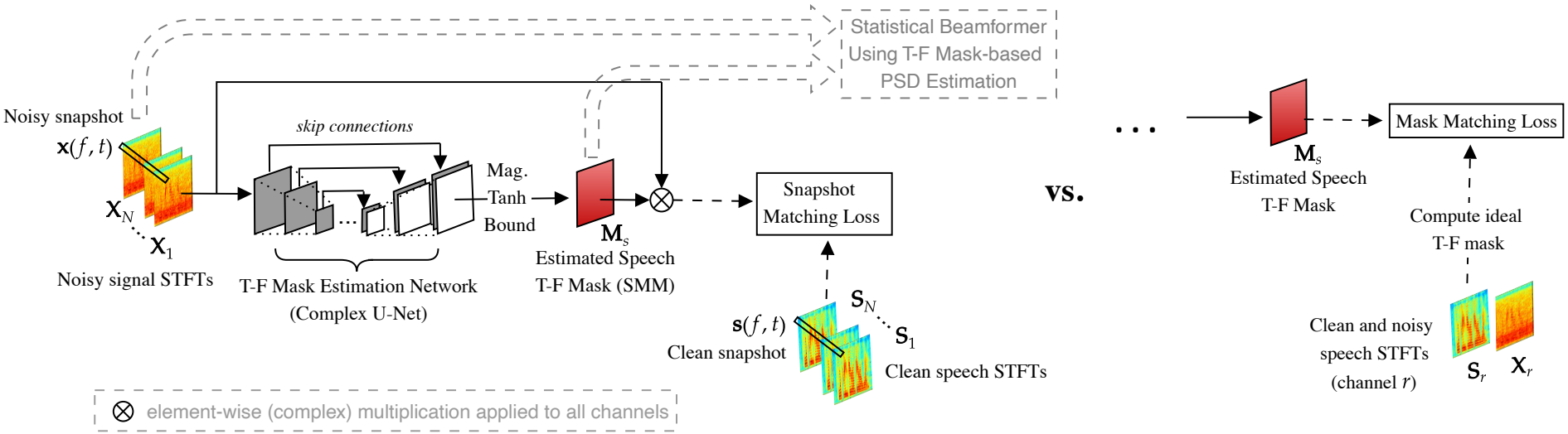

Figure: The left figure depicts the proposed SMM estimation network based on using a complex-valued U-Net architecture. The "Mag. Tanh Bound" module realizes the magnitude constraint within the unit circle. "Snapshot Matching Loss" computes the difference between the estimated and true signal snapshots based on combined power-law compressed MSE criterion.

Experimental Results

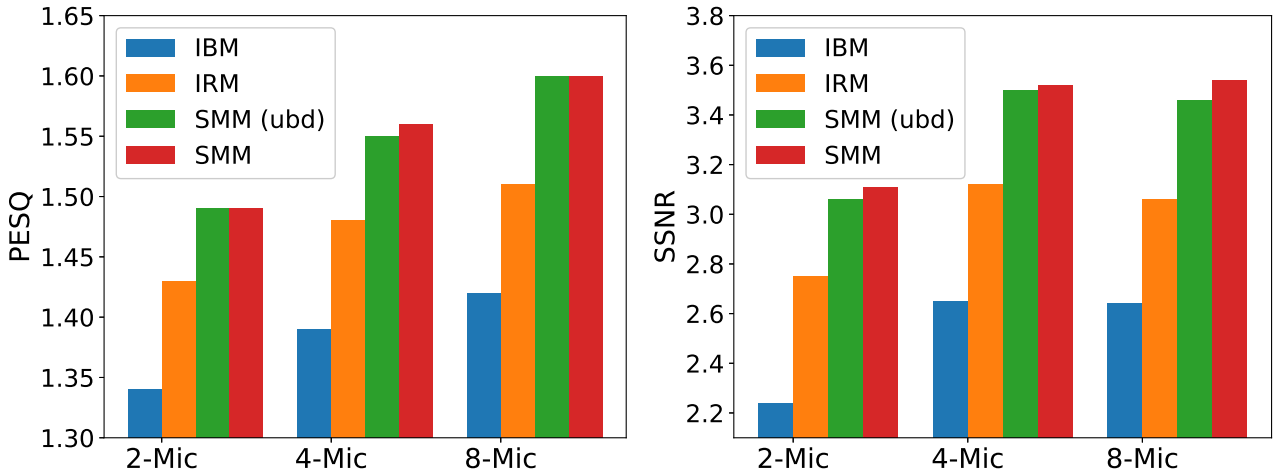

Figure: SE performance of T-F mask-based MWF using PSD matrices estimated from IBM, IRM, (ubd: unbounded) SMM, and SMM.

Table: Average Frobenius distance between the mask-based speech PSD estimate and the PSD estimated by using oracle speech snapshots for different masking schemes.

Estimated Mask Type	2-mic	4-mic	8-mic
IBM	48.88	144.60	185.81
IRM	42.31	133.57	169.19
SMM (ubd)	38.60	75.69	152.98
SMM	**38.07**	**75.65**	**147.52**

Experimental Results

Table: PESQ comparison of masking schemes on other datasets. "($\mathbb{R}$)" and "($\mathbb{C}$)" stand for using real or complex U-Net, respectively.

Dataset	Noisy	IBM($\mathbb{R}$)	IBM($\mathbb{C}$)	IRM($\mathbb{R}$)	IRM($\mathbb{C}$)	SMM
AVSpeech+Real RIRs	1.40	1.59	1.57	1.64	1.61	**1.73**
CHiME-3	1.27	1.83	1.85	2.18	2.18	**2.23**

Table: Comparison with existing deep learning-based SE methods.

Methods	Type	# Params	PESQ	STOI
Noisy	-	-	1.40	0.598
Conv-TasNet [1]	Single-channel	8.7M	1.64	0.638
DCUnet [2]	Single-channel	7.6M	1.62	0.631
FaSNet [3]	Multichannel	2.8M	1.71	0.652
SMM-based MWF (ours)	Multichannel	**1.3M**	**1.73**	**0.681**

[1] Luo and Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," in *IEEE/ACM TASLP*, 2019.

[2] Choi et al., "Phase-aware speech enhancement with deep complex U-Net," in *ICLR*, 2019.

[3] Luo et al., "End-to-end microphone permutation and number invariant multi-channel speech separation," in *ICASSP*, 2020.