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

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

Email: chinghua.l@samsung.com

Abstract

Objective:

- Develop a new masking strategy to improve time-frequency (T-F) mask-based neural beamforming algorithms for multichannel speech enhancement (SE)

Methods:

- Propose the **Snapshot Matching Mask (SMM)** that aims to minimize the distance between the predicted and the true signal snapshots, leading to *a more systematic way of estimating the speech and noise power spectral density (PSD) matrices* that are used to derive beamformer weights

Results

- SMM demonstrates improved SE performance compared to existing masking approaches (e.g., the ideal binary mask (IBM) and ideal ratio mask (IRM)) that *lack direct connection to PSD estimation* for mask-based neural beamforming

1 Background

1.1 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, the i-th microphone noisy signal STFT $\mathbf{X}_i \in \mathbb{C}^{F\times T}$ of an N-microphone array can be expressed as:

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

  $\forall f, t$, 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 $r \in \{1, \dots, N\}$ given the noisy $\mathbf{X}_1, \dots, \mathbf{X}_N$

1.2 T-F Mask-Based Neural Beamformer

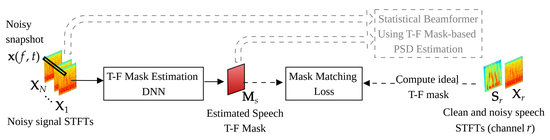

Figure 1: T-F mask-based neural beamformer: T-F mask estimation DNN (learning-based) followed by statistical beamformer (model-based).

- The T-F mask estimation DNN is utilized to predict some pre-defined 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 $\mathbf{s}(f,t) = [S_1(f,t), \dots, S_N(f,t)]^T$ and $\mathbf{v}(f,t) = [V_1(f,t), \dots, V_N(f,t)]^T$ are the speech and noise snapshots.

1.3 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, 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 masks is not based on multichannel characteristics but on single-channel SE solutions. As a result, they lack direct relation to PSD matrix estimation

2 Proposed Method

2.1 SMM Estimation Framework

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

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

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

Snapshot Matching Loss for optimizing (4):

$$\begin{aligned}\mathcal{L}(\hat{\mathbf{s}}(f,t), \mathbf{s}(f,t)) =& \frac{1}{N}\sum_{i=1}^{N} 0.7(|\hat{S}_i(f,t)|^{0.3} - |S_i(f,t)|^{0.3})^2 \\ &+0.3||\hat{S}_i(f,t)|^{0.3}e^{j\angle\hat{S}_i(f,t)} - |S_i(f,t)|^{0.3}e^{j\angle S_i(f,t)}|^2.\end{aligned} \tag{7}$$

PSD Updates with SMM:

$$\begin{aligned}\mathbf{\Phi}_s(f,t) =& \lambda_s\mathbf{\Phi}_s(f,t-1) + \hat{\mathbf{s}}(f,t)\hat{\mathbf{s}}^H(f,t) \\ =& \lambda_s\mathbf{\Phi}_s(f,t-1) + |M_s(f,t)|^2\mathbf{x}(f,t)\mathbf{x}^H(f,t).\end{aligned} \tag{8}$$

2.2 SMM Properties

- SMM simultaneously considers all channels together via a complex-valued masking scheme to directly minimize the distance between the estimated and the true signal snapshots.
- The magnitude constraint, i.e., $|M_s(f,t)| \le 1, \forall f, t$, leads to the value of $|M_s(f,t)|^2$ in the PSD update (2) lying in $[0,1]$ and can be interpreted as the speech presence probability (SPP) aligning with other T-F masks.
- Adopting the complex-valued mask for also manipulating the phase components of signal snapshots, SMM can better exploit spatial characteristics of multichannel signals, as compared to approaches like [1] that utilize real-valued T-F masks pre-defined on a single reference microphone.

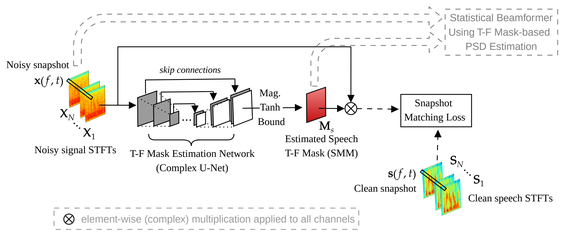

Figure 2: Proposed SMM estimation network based on using a complex-valued U-Net architecture. "Mag. Tanh Bound" realizes the magnitude constraint within the unit circle in (4). "Snapshot Matching Loss" computes the difference between the estimated and true signal snapshots based on (7).

3 Simulation Results

Table 1 compares the multichannel Wiener filter (MWF) beamformer outcomes of using the PSD matrices estimated based on oracle entities (to observe performance upper bound).

Table 1: SE performance of MWF using PSD estimated based on oracle IBM, IRM, and speech snapshots (SS). There is apparent performance gap between the results of using oracle IBM and IRM and the results of using oracle SS.

# Mic	PESQ			SSNR		
	Oracle IBM	Oracle IRM	Oracle SS	Oracle IBM	Oracle IRM	Oracle SS
2	1.42	1.68	**1.78**	2.02	4.47	**4.85**
4	1.51	1.85	**1.99**	2.42	5.25	**5.71**
8	1.87	2.05	**2.16**	4.04	5.56	**5.74**

Next, we evaluate the proposed SMM for better PSD estimation of T-F mask-based MWF in Figure 3. The SMM significantly outperforms IBM and IRM for settings. Such improvement could be attributed to the better PSD estimation of SMM over the IBM and IRM as revealed in Table 2.

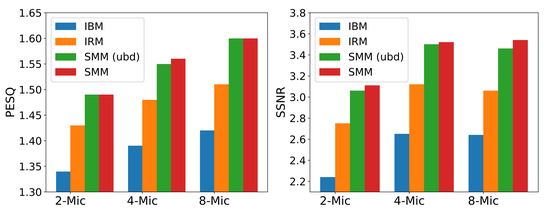

Figure 3: SE performance of T-F mask-based MWF using PSD matrices estimated from IBM, IRM, (ubd: unbounded) SMM, and SMM.

Table 2: Average Frobenius distance between the mask-based speech PSD estimate and the PSD estimated by using oracle speech snapshots for different masking schemes. One can see that the resulting distance of SMM-based PSD estimate is smallest as compared to the IBM- and IRM-based estimates.

# Mics	IBM	IRM	SMM (ubd)	SMM
2	48.88	42.31	38.60	**38.07**
4	144.60	133.57	75.69	**75.65**
8	185.81	169.19	152.98	**147.52**

We also present the estimated IRM and SMM for a noisy signal in Figure 4, evaluation on other two datasets in Table 3, and comparison with several existing deep learning-based SE methods in Table 4, to demonstrate the effectiveness of SMM.

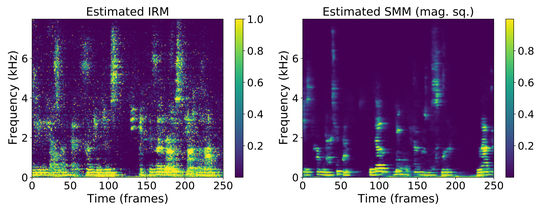

Figure 4: Estimated IRM and SMM of a noisy utterance. The estimated SMM shows a smoother distribution of the speech while the IRM is much noisier.

Table 3: PESQ comparison of masking schemes on other datasets. "(ℝ)" and "(ℂ)" stand for using real or complex U-Net, respectively.

Dataset	Noisy	IBM(ℝ)	IBM(ℂ)	IRM(ℝ)	IRM(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 4: Comparison with existing deep learning-based SE methods.

Methods	Type	# Params	PESQ	STOI
Noisy	-	-	1.40	0.598
Conv-TasNet [2]	Single-channel	8.7M	1.64	0.638
DCUnet [3]	Single-channel	7.6M	1.62	0.631
FaSNet [4]	Multichannel	2.8M	1.71	0.652
SMM-based MWF (ours)	Multichannel	**1.3M**	**1.73**	**0.681**

4 Conclusion

We proposed the SMM framework for improved PSD estimation in mask-based neural beamforming, which directly minimizes the distance between the estimated and true signal snapshots. Simulations show the potential of SMM to overcome the limitation of existing T-F masks that lack direct connection to PSD estimation.

References

[1] Chakrabarty and Habets, "Time-frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, 2019.

[2] Luo and Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2019.

[3] Choi et al., "Phase-aware speech enhancement with deep complex U-Net," in *Int. Conf. Learning Repres. (ICLR)*, 2019.

[4] Luo et al., "End-to-end microphone permutation and number invariant multi-channel speech separation," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020.