Better Exploiting Spatial Separability in Multichannel Speech Enhancement with an Align-and-Filter Network

Ching-Hua Lee, Chouchang Yang, Yashas Malur Saidutta, Rakshith Sharma Srinivasa, Yilin Shen, Hongxia Jin
AI Center-Mountian View, Samsung Electronics

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

Email: chinghua.l@samsung.com

Abstract

Objective:

- Develop a new deep learning-based multichannel speech enhancement (SE) model for improved robustness against spatially uncertain target speech scenarios

Methods:

- Propose the *Align-and-Filter Network (AFnet)* featuring a two-stage design inspired by the *alignment-followed-by-filtering* principle from classical signal processing

Results

- By leveraging *relative transfer functions (RTFs)* as the training target for spatial alignment, AFnet learns interpretable directional features to better exploit spatial separability of sound sources for improved SE performance

1 Overview

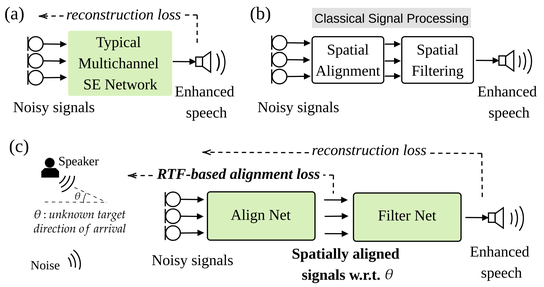

Figure 1: Illustration of different multichannel SE systems. (a) Typical deep learning-based methods directly model the noisy-to-clean speech mapping. (b) Classical signal processing approaches perform SE as two sub-tasks. (c) Our framework models both alignment and filtering processes with deep networks, achieved by supervising alignment based on relative transfer functions (RTFs).

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:** Time-frequency 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:

$$\mathbf{X}_i = \mathbf{H}_i \odot \mathbf{S}_0 + \mathbf{V}_i = \mathbf{S}_i + \mathbf{V}_i, \tag{1}$$

where $\mathbf{S}_i \triangleq \mathbf{H}_i \odot \mathbf{S}_0 \in \mathbb{C}^{F\times T}$ is the speech component received by microphone i, $\odot$ denotes element-wise product, $\mathbf{H}_i \in \mathbb{C}^{F\times T}$ is the acoustic transfer function between the speech source $\mathbf{S}_0 \in \mathbb{C}^{F\times T}$ and microphone i, and $\mathbf{V}_i \in \mathbb{C}^{F\times T}$ is the noise component captured by microphone i.

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

2.2 Beamforming in STFT Domain

Multichannel SE systems usually perform spatial filtering, or beamformig, through proper *linear combination* of the microphone signals to obtain the enhanced signal $\hat{\mathbf{S}} \in \mathbb{C}^{F\times T}$:

$$\hat{\mathbf{S}} = \sum_{i=1}^{N} \mathbf{W}_i \odot \mathbf{X}_i, \tag{2}$$

where $\mathbf{W}_i \in \mathbb{C}^{F\times T}$ is the set of beamformer filters of mic i. To derive $\mathbf{W}_i$'s, traditional signal processing algorithms typically employ a spatial alignment stage by estimating the **temporal (phase)** and **level (magnitude)** differences among the received speech components $\mathbf{S}_i$'s. *However, the notion of such spatial alignment is often overlooked in the SE deep network design.*

3 Proposed Method

3.1 AFnet

- Our AFnet interprets multichannel SE process as:

$$\hat{\mathbf{S}} = \sum_{i=1}^{N} \underbrace{(\mathbf{F}_i \odot \mathbf{A}_i)}_{\text{beamformer } \mathbf{W}_i} \odot \mathbf{X}_i = \sum_{i=1}^{N} \mathbf{F}_i \odot \underbrace{(\mathbf{A}_i \odot \mathbf{X}_i)}_{\text{aligned signals } \mathbf{Z}_i}. \tag{3}$$

This suggests that the beamformer weights $\mathbf{W}_i$ in (2) be decomposed into *spatial alignment ($\mathbf{A}_i$) and filtering ($\mathbf{F}_i$) units*.

- The process is realized by the *sequential masking* scheme using two deep net modules, Align Net and Filter Net, as depicted in Figure 2:

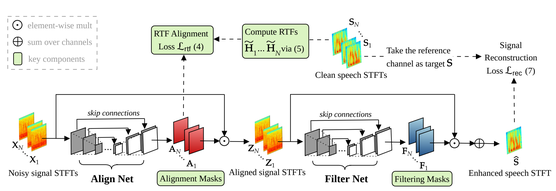

Figure 2: AFnet based on the "*align-then-filter*" principle for better capturing spatial characteristics of speech data rich in directional variety. We highlight (in green) the key components that contribute to our SE improvement: **RTF-based supervision for spatial alignment** and **sequential masking** design.

3.2 Learning the Alignment Process

- We propose the optimization for training the Align Net by:

$$\min \quad \mathcal{L}_{\text{rtf}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{A}_i - \tilde{\mathbf{H}}_i\|_F^2, \tag{4}$$

where the alignment mask $\mathbf{A}_i$ estimates the **speech RTF** $\tilde{\mathbf{H}}_i$ defined between microphone i and reference microphone r:

$$\tilde{\mathbf{H}}_i \triangleq \mathbf{H}_r \oslash \mathbf{H}_i = (\mathbf{H}_r \odot \mathbf{S}_0) \oslash (\mathbf{H}_i \odot \mathbf{S}_0) = \mathbf{S}_r \oslash \mathbf{S}_i. \tag{5}$$

- Ideally, if $\mathbf{A}_i = \tilde{\mathbf{H}}_i$ we get the perfectly aligned signal as:

$$\mathbf{Z}_i = \tilde{\mathbf{H}}_i \odot \mathbf{X}_i = (\mathbf{S}_r \oslash \mathbf{S}_i) \odot (\mathbf{S}_i + \mathbf{V}_i) = \mathbf{S}_r + \tilde{\mathbf{V}}_i, \tag{6}$$

where each $\mathbf{Z}_i$ contains the same speech component $\mathbf{S}_r$ independent of the microphone index i.

3.3 Learning the Filtering Process

- As the goal is to reconstruct the clean speech at the final output, we train the entire AFnet (i.e., Align Net + Filter Net) by minimizing the reconstruction loss between $\mathbf{S}$ and $\hat{\mathbf{S}}$:

$$\min \quad \mathcal{L}_{\text{rec}} = 0.3\|\hat{\mathbf{S}}^{0.3} - \mathbf{S}^{0.3}\|_F^2 + 0.7\|\|\hat{\mathbf{S}}|^{0.3} - |\mathbf{S}|^{0.3}\|_F^2. \tag{7}$$

4 Simulation Results

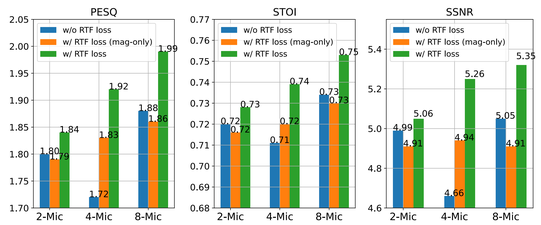

Figure 3: SE comparison of different RTF alignment schemes for AFnet training. We see that *"w/ RTF loss"* achieves considerable improvements in all settings, suggesting the advantages of performing spatial alignment.

Figure 4: t-SNE of learned alignment masks for signals coming from three different locations (loc 1,2,3). AFnet trained with RTF alignment loss results in separate clusters, corresponding to successively learned spatial separability.

Table 1: Generalization to unseen and dynamic acoustic environments.

Method	Unseen rooms			Time-varying location		
	PESQ	STOI	SSNR	PESQ	STOI	SSNR
AFnet w/o RTF loss	1.77	0.712	4.93	1.71	0.700	4.52
AFnet	**1.95**	**0.740**	**5.47**	**1.82**	**0.720**	**5.00**

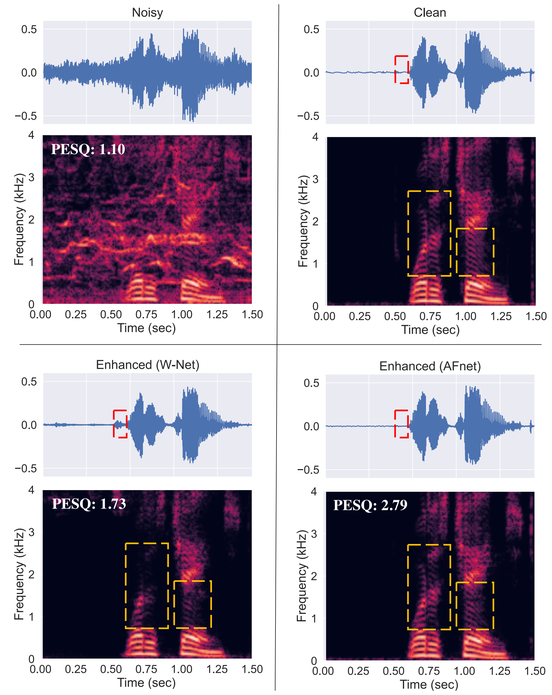

Figure 5: Comparing the waveforms of W-Net (an alignment-unaware model representing Figure 1 (a)) and AFnet (8-mic), we can see that AFnet successfully removes the noise-only segment in the beginning of the utterance while W-Net fails to, as indicated in the red boxes. By inspecting the spectrograms, we notice that some detailed speech structures are highly distorted with W-Net, while AFnet preserves more speech structures, as marked by yellow boxes.

5 Conclusion

We presented AFnet, a multichannel SE deep learning framework that exploits the "*align-then-filter*" notion to handle speech sources with spatial uncertainty by leveraging the RTF, an essential component in many signal processing-based algorithms. Our findings suggest that alignment indeed plays an important role in deep learning-based approaches, especially for spatially diverse speech scenarios.