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

Artificial Intelligence Center - Mountain View, Samsung Electronics

The 50th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) @ Hyderabad, India – April 2025

Introduction

- Most of existing deep learning-based multichannel speech enhancement (SE) approaches directly model the noisy-to-clean speech mapping in *one-stage*, lacking explicit spatial feature learning which in turn leads to *reduced robustness against uncertainty of target locations.*

- On the other hand, conventional signal processing-based methods usually adopt a *two-stage* design: **spatial alignment** followed by **noise filtering**, mitigating the uncertainty of target speech locations via explicitly aligning the target speech based on spatial information.

- In this work, we propose *Align-and-Filter Network (AFnet)*, a deep learning framework featuring a two-stage design inspired by the *alignment-followed-by-filtering* principle from classical signal processing, to improve the robustness against spatial uncertainty of target speech in deep learning-based methods.

Overview of Methods

- Typical deep learning-based methods directly model the noisy-to-clean speech mapping – Figure (a)
- Signal processing approaches perform SE as two sub-tasks – Figure (b)
- Our framework models both alignment and filtering processes with deep networks, by supervising the spatial alignment using **relative transfer functions (RTFs)** as the training target – Figure (c)

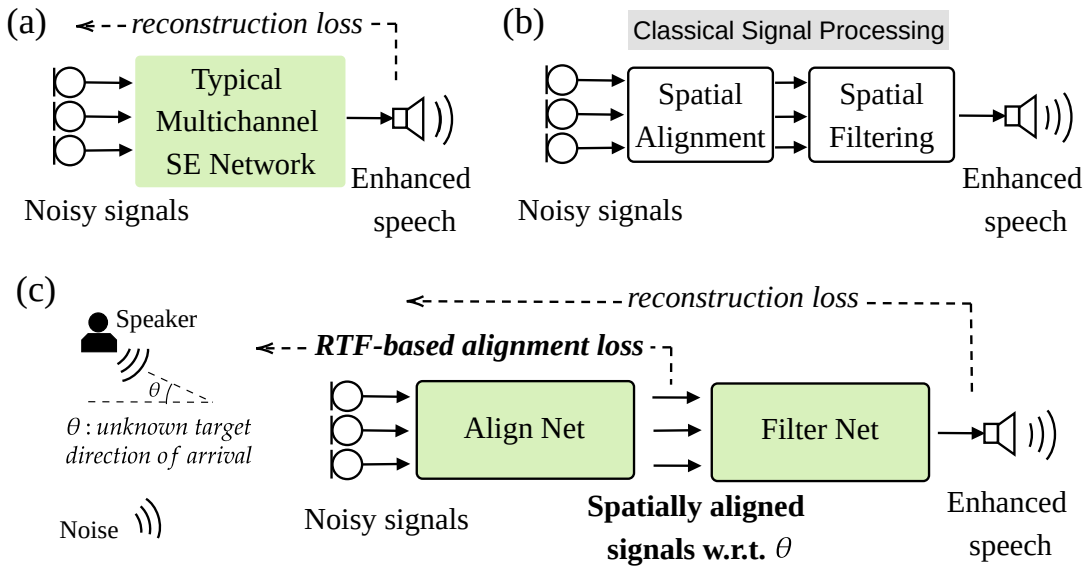

Figure: Illustration of different multichannel SE systems.

Problem Formulation

- **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:

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:

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

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 $\mathbf{S}_r \in \mathbb{C}^{F\times T}$ of a reference microphone $r \in \{1, \ldots, N\}$ given the N noisy signals $\mathbf{X}_1, \ldots, \mathbf{X}_N$

Spatial Filtering in STFT Domain

- In the 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 microphone i.

- When deriving $\mathbf{W}_i$'s, traditional signal processing algorithms typically require a spatial alignment stage by estimating the **temporal (phase)** and **level (magnitude)** differences among the received speech components $\mathbf{S}_i$'s at all microphones.

- In many deep learning-based multichannel SE systems, the notion of such spatial alignment is often overlooked in the SE network design. In this work, we advocate for the importance of the alignment process by explicitly integrating it into the learning framework.

Proposed Method

- Our AFnet interprets multichannel SE process as:

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

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

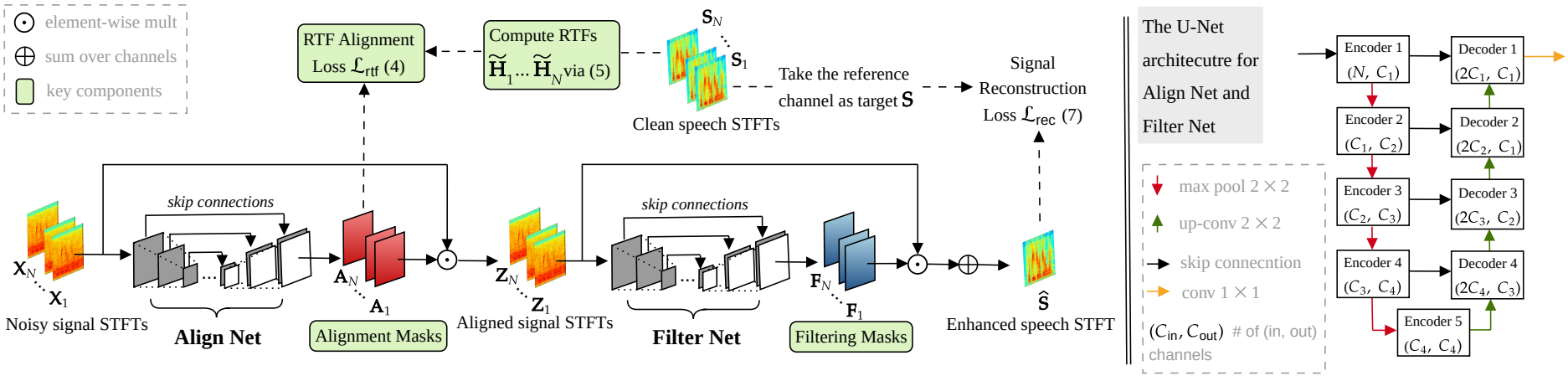

Figure: Proposed AFnet based on the "*align-then-filter*" principle for better capturing spatial characteristics of speech data rich in directional variety.

Proposed Method

- **Learning the Alignment Process**: We propose using the speech RTFs as the training target for the Align Net by:

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

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 aligned signal as (by using (1) and (5)):

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

- **Learning the Filtering Process**: As the goal is to reconstruct the clean speech at the final output, we train the entire AFnet model by:

$$\min \quad \mathcal{L}_{\text{rec}} = \beta\|\hat{\mathbf{S}}^c - \mathbf{S}^c\|_F^2 + (1-\beta)\||\hat{\mathbf{S}}|^c - |\mathbf{S}|^c\|_F^2, \tag{7}$$

In this work we use $\beta = 0.3$ and $c = 0.3$ for the reconstruction loss.

Experimental Results

Table: SE performance of aligned and unaligned input signals.

# Mic	PESQ		STOI		SSNR	
	Aligned	Unaligned	Aligned	Unaligned	Aligned	Unaligned
2	**2.68**	1.68	**0.842**	0.697	**7.86**	4.66
4	**2.89**	1.67	**0.862**	0.696	**8.28**	4.56
8	**3.04**	1.76	**0.879**	0.719	**8.66**	4.68

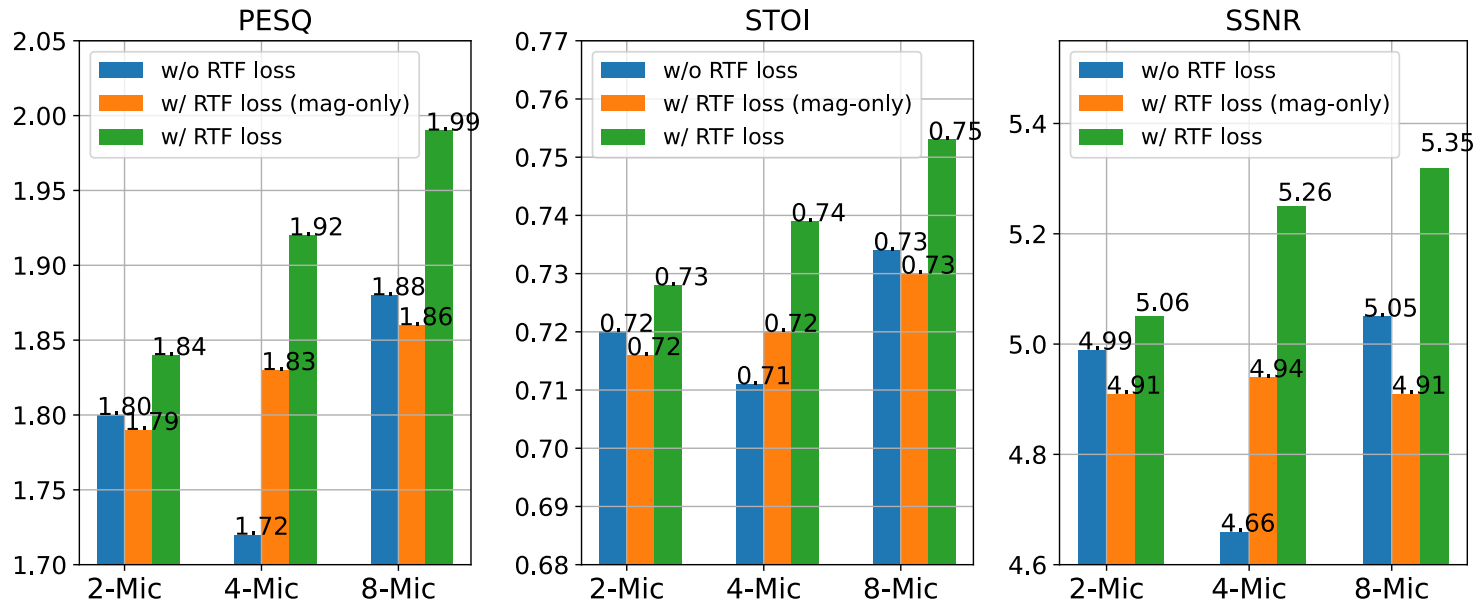

Figure: SE comparison of different RTF alignment schemes for AFnet training.

Experimental Results

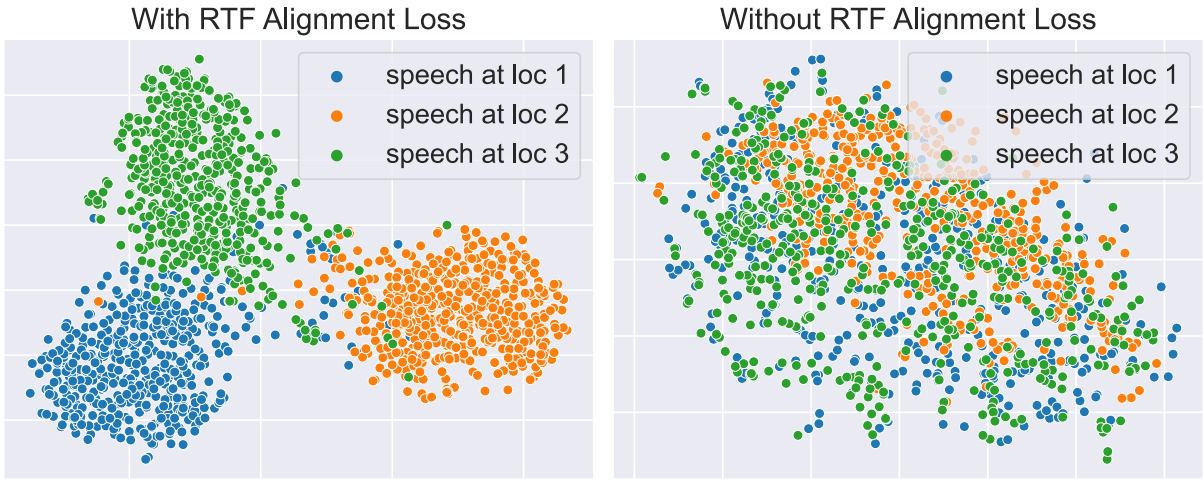

Figure: t-SNE of learned alignment masks for signals coming from three different locations (loc 1, loc 2, loc 3). AFnet trained with RTF alignment loss supervision results in separate clusters, corresponding to successively learned spatial separability.

Table: 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**

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. We showed that utilizing RTFs as the training target is an effective way for the model to learn the alignment process.
- More broadly, our work suggests that it would be beneficial to consider such spatial alignment aspect when developing advanced multichannel SE systems.