EXPLOITING SPATIAL SEPARABILITY FOR MULTI-CHANNEL SPEECH ENHANCEMENT WITH AN ALIGN-AND-FILTER FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Multichannel speech enhancement (SE) systems separate the target speech from background noise by performing spatial and spectral filtering. The development of multichannel SE has a long history in the signal processing field, where one crucial step is to leverage spatial separability of sound sources by aligning the microphone signals in response to the target speech source prior to further filtering processes. However, most existing deep learning based multichannel SE works have yet to effectively incorporate or emphasize this spatial alignment aspect in the network design – we postulate that it is owing to the lack of suitable datasets with sufficient spatial diversity of the sound sources. In this paper, we highlight this important but often overlooked step in deep learning based multichannel SE, i.e., signal alignment, by introducing a novel Align-and-Filter framework featuring a two-stage design. Under this framework, the Align-and-Filter network (AFnet) is proposed, which exploits the relative transfer functions (RTFs) for aligning signals. For exploration purposes, we argue that the popular CHiME-3 multichannel dataset has its own limitation in representing spatially diverse speech signals, and adopt simulated and real-world measured room impulse responses to generate multichannel recordings where the target speech sources might come from arbitrary directions. Our findings suggest that for spatially diverse acoustic settings, utilizing the RTFs for alignment purposes in the two-stage framework significantly improves the capability of the network to separate the target speech from the noise signals as exploiting their spatial separability. More broadly, our studies advocate for the advantages and significance of considering the signal alignment aspect, a wisdom coming from conventional signal processing, for developing future deep based multichannel SE algorithms to improve enhancement outcomes with positional diverse target speech scenarios.

1 INTRODUCTION

Signal quality could significantly deteriorate in the presence of noise and interference. Information carried by the signal, e.g., speech, could be severely lost at the receiver side (the human ear or smart assistants) in a very noisy environment, causing difficulty of communication. Speech enhancement (SE) techniques aim to resolve such issues, which had been majorly developed based on model-based, signal processing approaches that utilize mathematical modeling of signal statistics Loizou (2007). More recently, the advent of deep learning has demonstrated the potential of advancing the development of SE systems to a higher level, making SE a topic of great interest in the machine learning field Wang & Chen (2018); Haeb-Umbach et al. (2019).

SE systems can be categorized into *single-channel* (single microphone) and *multichannel* (multiple microphones) schemes. An important aspect of multichannel SE against single-channel SE is the exploitation of *spatial separability*, as known as *spatial filtering* or *beamforming*, enabled by the *difference* between the amplitudes and times of arrival of the received microphone signals due to the different acoustics paths the sound waveform travels to the microphones. In many conventional signal processing methods for beamforming, a key step is to *align* the microphone signals in response to the target signal source before any further filtering processes. This step, by *steering* the array toward the location of target signal, aims to compensate for the difference of the amplitudes and time delays (or correspondingly the magnitudes and phases in the frequency domain) of the microphone

signals with respect to the target source. Ideally, after the alignment step, each microphone should contain the same target speech component with no difference in amplitude and time delay (or magnitude and phase). That is, perfectly steering the microphone array to the target signal makes it as if the target signal comes from the broadside (i.e., front side). Thus, an efficient SE system can first align its microphone signals in response to the target speech, followed by spectral processing for fine-tuning and enhancement. This is similar to the human ear which can do both delay estimation and frequency analysis. Surprisingly, though well-known in many conventional signal processing approaches, there are only few deep learning SE systems designed based on this observation.

In this paper, we revisit this important alignment step in conventional signal processing approaches for the deep learning based multichannel SE schemes. The main contributions of this paper are summarized as follows:

- We propose a new deep learning framework, namely Align-and-Filter network (AFnet) (Figure 1), to study the spatial separability in multichannel SE. The new framework features a two-stage design, i.e., Align Net and Filter Net, enabling the incorporation of spatial information to perform signal alignment for improved enhancement outcomes.
- A novel training protocol is presented which leverages the *relative transfer function (RTF)* information for facilitating signal alignment during the learning stage. We show that by leveraging RTF information the network can more effectively learn the spatial characteristics of the target speech, thereby exploiting spatial separability in a more systematic way for denoising purposes.
- The proposed AFnet framework is utilized to analyze the spatial diversity of speech sources in different datasets. We point out that some existing dataset such as CHiME-3 lacks sufficient positional variety of target speech and thus has its own limitation of representing broader SE scenarios.
- On more spatially diverse data, our results indicate that significant improvements can be achieved by exploiting RTFs in the network learning. We therefore highlight the wisdom of conventional signal processing methods – *the importance of signal alignment* – and advocate for incorporating the concept into the development of deep learning based multichannel SE systems, especially when the target speech may locate in arbitrary positions.

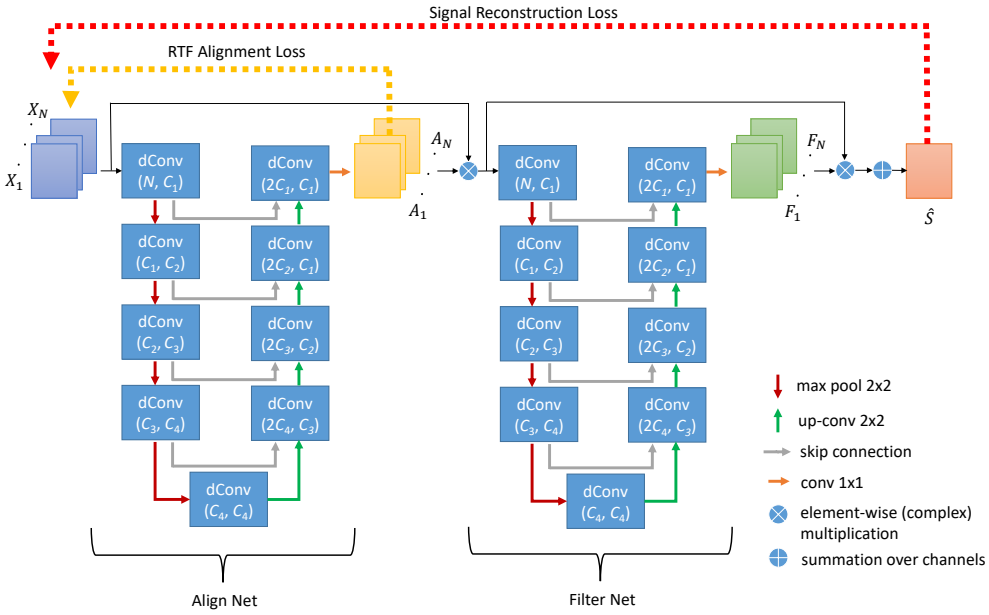

Figure 1: The proposed AFnet utilizing a (complex) W-Net architecture. The first half of the AFnet is a single (complex) U-Net that takes the microphone signals as input and estimates a set of alignment masks. The second half is another (complex) U-Net that takes the aligned signals as input and estimate a set of filtering masks for target signal enhancement.

2 RELATED WORK

Multichannel SE has been a well known topic in signal processing for decades. Aside from leveraging temporal and spectral characteristics, multichannel SE can exploit positional information to perform *spatial filtering*, or *beamforming*, that allows extracting the target speech from noise based on spatial separability. Conventional beamforming approaches rely on the so-called "steering vector" which carries positional information about the target speech Doclo et al. (2015); Trees (2004), e.g., the minimum-variance-distortionless-response (MVDR) beamformer Capon (1969) and its variants Frost (1972); Griffiths & Jim (1982). It is an important step in beamforming that the microphone array is steered toward the target signal location prior to further filtering processes.

To steer the array toward the target signal, it is typically required that the acoustic paths between the target signal and the microphones be known, often characterized as frequency domain acoustic transfer functions. Due to the challenge of estimating the absolute paths, many multichannel SE systems instead identify the ratio of the acoustic transfer functions, i.e., the *relative transfer function (RTF)*, that represents the coupling between sensors in response to a desired source Cohen (2004). Many advanced signal processing SE approaches have exploited the information inherent in RTFs Gannot et al. (2001); Krueger et al. (2010); Koldovskỳ et al. (2015).

More recently, supervised learning approaches, especially deep learning, have significantly changed the way of developing SE systems. Along with the success of deep neural networks (DNNs) on single-channel SE, several multichannel SE systems have also been proposed, e.g., Erdogan et al. (2016); Bu et al. (2019); Liu et al. (2020); Luo et al. (2020); Tolooshams et al. (2020); Koldovskỳ et al. (2015). However, the utilization of such important RTF information is often overlooked in the model design of DNN-based multichannel SE. To the best of our knowledge, the only work that incorporates RTFs for SE is Wang & Wang (2018), in which, however, the RTF estimation is used as an intermediate step to assist the conventional MVDR beamformer only. We postulate that the overlook may be due to the lack of sufficient spatial variety of the speech sources in the popular dataset such as CHiME-3, where the benefit of utilizing RTF information could be only marginal. However, many practical situations can have the target speech coming from arbitrary directions and thus a deep dive into the RTF spatial alignment aspect is still of great importance.

3 PROPOSED METHOD

We consider an acoustic scenario with one desired speech source and several interfering noise signals in a reverberant environment. The SE system will be developed in the time-frequency domain using the short-time Fourier transform (STFT) Parchami et al. (2016) where the time-domain waveforms are transformed to complex-valued STFT representations. Let f, t stand for the frequency index and time frame index, we consider an additive noise model where the i-th microphone noisy signal $X_i(f,t)$ of an N-element local microphone array can be written in the STFT domain as Doclo et al. (2015): $X_i(f,t) = S_i(f,t) + V_i(f,t)$, with $S_i(f,t) = H_i(f,t)S_0(f,t)$ being the speech component received by microphone i, where $H_i(f,t)$ is the acoustic transfer function between microphone i and the location of the far-field speech source signal $S_0(f,t)$, and $V_i(f,t)$ is the noise component captured by microphone i. Typically, the goal of multichannel SE is to recover the speech component $S_r(f,t)$ of a selected reference microphone $r \in \{1, \ldots, N\}$ given the noisy microphone signals $X_1(f,t), \ldots, X_N(f,t)$. Sometimes it may be required to recover the far-field signal $S_0(f,t)$ depending on the applications.

Multichannel SE systems typically perform the "filter-and-sum" operation, or more generally known as "beamforming" – linearly combining the signals of different microphone channels to extract the target signal S from background noise, where S can be S_r of the reference microphone or even the far-field source S_0. In the time-frequency domain, the estimation process can be expressed as: $\hat{S}(f,t) = \sum_{i=1}^{N} W_i(f,t)X_i(f,t)$, where $W_i(f,t)$ is the corresponding filter weight for microphone i and $\hat{S}(f,t)$ is the enhanced signal. Conventional signal processing based beamforming approaches utilize signal and noise statistics to come up with a set of proper filter weights for extracting the target source. Recent deep learning based approaches rely on DNNs to estimate the filter weights, where $X_i, W_i, \hat{S}_i \in \mathbb{C}^{F \times T}$ are the STFT representations for the i-th microphone (F and T stand for the number of frequency bins and the number of time frames, respectively).

3.1 ALIGN-AND-FILTER FRAMEWORK

Inspired by the alignment concept in signal processing based algorithms, we propose a novel Align-and-Filter framework for DNN-based multichannel SE as illustrated in Figure 1. It features a two-stage design, where two blocks of network are utilized to carry out alignment and filtering purposes respectively. The first stage (Align Net) aims at estimating a set *alignment masks* for steering the input microphone signals toward the desired signal location. The second stage (Filter Net) then focuses on cleaning up the interference in the aligned signals by estimating another set of *filtering masks* and linearly combines all channel outputs into the enhanced signal. Overall, the enhancement process can be expressed as: $\hat{S}(f,t) = \sum_{i=1}^{N} F_i(f,t)A_i(f,t)X_i(f,t)$ where $A_i \in \mathbb{C}^{F\times T}$ is the set of alignment masks and $F_i \in \mathbb{C}^{F\times T}$ is the set of filtering masks of the i-th microphone.

3.2 RTF-AWARE TRAINING

The spatial relationship between the far-field target source and the microphones along with the acoustic environment characteristics determine the inter-channel level difference (ILD) and inter-channel time difference (ITD) of the microphone signals, which are important cues for sound source localization. In the STFT domain, the ILD and ITD correspond to the magnitude and phase of the RTF respectively, which is the ratio between the acoustic transfer functions of two microphone channels with regards to the sound source Li et al. (2016). The RTFs $\tilde{H}_i$ can be expressed as:

$$\tilde{H}_i(f,t) \triangleq \frac{H_r(f,t)}{H_i(f,t)} = \frac{H_r(f,t)S_0(f,t)}{H_i(f,t)S_0(f,t)} = \frac{S_r(f,t)}{S_i(f,t)}, \tag{1}$$

$\forall i = 1,\ldots,N$, where $H_r(f,t)$ denotes the acoustic transfer function of the selected reference microphone $r \in \{1,\ldots,N\}$. From (1) we can see that the RTFs of the target speech can be obtained from the ratios of the clean microphone signals.

Channel-wise multiplying the RTF $\tilde{H}_i(f,t)$ with the microphone signal $X_i(f,t)$ we obtain:

$$Z_i(f,t) = \tilde{H}_i(f,t)X_i(f,t) = \frac{H_r(f,t)}{H_i(f,t)}[H_i(f,t)S_0(f,t) + V_i(f,t)] = S_r(f,t) + \tilde{V}_i(f,t), \tag{2}$$

$\forall i = 1,\ldots,N$, where $S_r(f,t)$ is the speech component of the reference microphone and $\tilde{V}_i(f,t) = [H_r(f,t)/H_i(f,t)]V_i(f,t)$ is the noise component. It can be seen that (element-wise) multiplying the RTFs with the microphone signals leads to aligning all the microphone channels with respect to the clean speech component at the reference microphone as each $Z_i(f,t)$ contains $S_r(f,t)$. In the case of far-field, anechoic setting, this means the microphone array has steered to the direction of the signal source. Equivalently, it can be viewed as the signal is coming from the broadside (front side) as each microphone should receive the same speech component when the acoustic paths to the microphones are the same.

For end-to-end training under the Align-and-Filter framework while utilizing the RTF information, we implicitly incorporate RTFs during the model learning phase via a novel RTF-aware training protocol consisting of two stages: first adjusting only the Align Net parameters so that the alignment masks A_i approximate the RTFs $\tilde{H}_i$ by minimizing the following RTF alignment loss: $\mathcal{L}_{rtf} = \sum_{i,f,t}|A_i(f,t) - \tilde{H}_i(f,t)|^2$, where we have chosen to use the mean squared error (MSE) loss. Optimizing for this loss function guides the network to learn to steer the microphone signals toward the target speech, making the model aware of the spatial information of the target source. The second stage performs parameter learning of the whole model (i.e., jointly adjusting both Align Net and Filter Net parameters) to reconstruct the clean speech. To this end, we minimize the signal reconstruction loss $\mathcal{L}_{rec}$ of the combined power-law compressed MSE loss in Wilson et al. (2018); Braun & Tashev (2021).

3.3 AFNET BASED ON MODIFIED W-NET ARCHITECTURE

We introduce the Align-and-Filter network (AFnet) for multichannel SE based on implementing a W-Net architecture suitable for realizing the two-stage framework as depicted in Figure 1. The W-Net consists of two U-Net blocks, where the first U-Net block serves as the Align Net and the second as the Filter Net. Originally, the W-Net Xia & Kulis (2017) is an improved version of the U-Net

structure which is a well known architecture consisting of a convolutional autoencoder with skip-connections for medical imaging Ronneberger et al. (2015). Both the original W-Net and U-Net are real-valued networks as it only requires to process pixels of images which are also real-valued. For the SE application here, extension to complex operations can be more suitable when working in the STFT domain. In Choi et al. (2018), the authors have proposed to utilize complex-valued U-Net for the single-channel SE task with several modifications. In this paper, we follow this idea and extend the W-Net by incorporating complex operations for the multichannel SE task.

Main modifications to the W-Net include: Convolutional layers of W-Net are all replaced to complex convolutional layers. For the activation function, complex leaky ReLU, i.e., an activation function which applies leaky ReLU on both real and imaginary values, are utilized. The number of feature maps in each layer are also modified. Note that in Figure 1, "dConv (C_{in}, C_{out})" stands for "double Convolutions" with C_{in} input channels and C_{out} output channels. Each dConv unit consists of two stacks of "(complex) 3×3 convolution→batch normalization→leaky ReLU". The convolution layer of the first stack takes C_{in} channels and outputs C_{out} channels; the convolution layer of the second takes C_{out} channels and outputs C_{out} channels.

4 EXPERIMENTS

4.1 DATASETS AND EXPERIMENTAL SETUP

4.1.1 PRIMARY DATASET

Speech corpus: We leverage the AVSpeech dataset which is a large collection of video clips of single speakers talking with no audio background interference introduced in Ephrat et al. (2018). The dataset is based on public instructional YouTube videos, from which short, 3-10 second clips were automatically extracted, where the only audible sound in the soundtrack belongs to a single speaking person. We downloaded 8308 clips from the training set and another 1199 clips from the test set for our experiments, using only the audio portion (converted to 16 kHz .wav files).

Room impulse response data: We mix the speech files with noise files of several types also downloaded from YouTube and consider different reverberation conditions for generating the noisy data using simulated impulse responses via Pyroomacoustics Scheibler et al. (2018), which is a Python-based acoustic simulator for generating different acoustic scenes. We create a room of size $8 \times 8 \times 3$ (length×width×height in meters). Reverberation time is randomly chosen from {0.16, 0.32, 0.48, 0.64} second. The simulating approach for estimating room impulse responses is based on the image source method. Speech and noise signals are randomly positioned at a distance of {1, 2, 3, 4} meters from the microphone array. 4 types of commonly seen interference (blender, vacuum, washer, baby cry) are used for training and another 4 types (dog barking, kids playing sound, hair dryer, food sizzling) are used for testing. Each type of noise is multiplied with a scale randomly chosen from {0, 0.5, 1, 1.5, 2} before added up, and the combined noise is then added to the clean speech according to a specific SNR level randomly selected from {-10, -6, -3, 0, 3, 6, 10} dB. In this way, we are adopting the supervised learning SE framework of mix-and-separate scheme, where the clean and noise signals have to be collected in advance and then mixed together to become the noisy signals for training purposes. This mix-and-separate scheme by assuming additive noise model is still being used in the majority of SE works (not only the classical model-based approaches but also recent data-driven methods). To apply to real data, we will collect the clean and noise recordings using the target device in real-world environments and then mix them up to generate the noisy mixtures for training purposes.

The main reason why we have chosen to use the simulated multichannel data as the primary dataset for exploration is for having sufficient spatial diversity of the target speech and also be able to specify the directions of arrivals for visualization purposes of the intermediate variables learned by the AFnet (i.e., the alignment masks) and other in-depth study (e.g., the effect of the number of microphones). The acoustic room simulator provides the flexibility of modifying several factors for generating representative multichannel data, which has also been a common practice in the majority of multichannel audio processing works. To further validate the applicability to more realistic scenarios, we also consider generating multichannel audio data using real-world measured acoustic room impulse responses from Hadad et al. (2014) with the results presented in Appendix.

4.1.2 AUXILIARY DATASET:

The publicly available CHiME-3 dataset Barker et al. (2015), made available as part of a speech separation and recognition challenge, is used for training and evaluating SE performance. The dataset is a 6-channel microphone recording of talkers speaking in a noisy environment, sampled at 16 kHz. It consists of 7,138 and 1,320 simulated utterances with an average length of 3 seconds for training and testing, respectively. We take this dataset only as auxiliary due to limited positional diversity of the target speech. Moreover, the dataset does not provide the ground-truth position of the target speech source. As such, the dataset is not ideal for exploring spatial separability purposes.

4.2 MODEL SETTINGS AND EVALUATION METRICS

In this paper, we set the number of channels of the AFnet (Figure 1) as: $C_1, C_2, C_3, C_4 = 32, 64, 64, 64$, resulting in a model size of approximately 2.6M parameters. For the STFT processing of the input signal, for the primary dataset we use the Hann window with 512-point fast Fourier transform (FFT) and for CHiME-3 with 1024-point FFT, both using a hop size of 256. During training, 4-second long segments are randomly cropped from the training data samples while during testing the whole utterances are used. The RTFs are obtained by computing the ratios between the clean speech signals at all microphones according to (1). Note that we truncate the magnitude of the RTFs to a maximum value of 10 for avoiding training instability.

Table 1: Speech enhancement performance comparison between aligned (A) and unaligned (U) input signals. Enhancement results for aligned signals are much better than unaligned signals, indicating the importance of signal alignment.

# Mic	PESQ		STOI		SSNR	
	A	U	A	U	A	U
2	2.04	1.37	0.783	0.614	5.28	2.27
4	2.17	1.47	0.806	0.654	6.05	3.01
8	2.33	1.48	0.820	0.655	6.45	2.97
CHiME-3	2.62	2.42	0.973	0.963	10.32	9.13

We use the following three commonly seen speech enhancement metrics for evaluation purposes. **PESQ:** Perceptual Evaluation of Speech Quality ITU-T Recommendation P.862.2 (2005) is a speech quality measure using the wide-band version recommended in ITU-T P.862.2 (value from -0.5 to 4.5, the higher the better). **STOI:** Short-Time Objective Intelligibility Taal et al. (2011) is a function that well represents the average intelligibility of the degraded speech. It provides a value from 0 to 1, with higher values indicating better intelligibility. **SSNR:** Segmental Signal-to-Noise Ratio Hansen & Pellom (1998) is an SNR measure, instead of working on the whole signal, that calculates the average of the SNR values of short segments.

4.3 RESULTS

4.3.1 ALIGNED VS. UNALIGNED SIGNAL ENHANCEMENT

To get an understanding of the importance of alignment, we first conduct an experiment to compare the difference between enhancement outcomes of aligned and unaligned signals. To this end, we manually align the input noisy microphone signals by multiplying them with the corresponding ground truth RTFs. This step basically replaces the function of an ideal Align Net. The aligned signals are directly fed into the Filter Net (the second half of AFnet) for filtering purposes to obtain the enhanced signal. For comparison, we also feed the original unaligned signals directly into the Filter Net and obtain the enhanced speech. The performance numbers are presented in Table 1. It can be seen that when the signals are first aligned, the Filter Net can do a much better job of separating the target signal from background noise and achieves significantly better scores than unaligned results in all three metrics. This observation is consistent across all three number of microphone settings. It is also worth noting that more improvement is observed for the aligned case when the number of microphone increases, while not necessarily or only limited improvement for the unaligned case. This indicates that signal alignment plays a more important role for a SE system with more number of microphones. The results from Table 1 suggest that incorporating the RTF information will benefit multichannel SE especially when there are more microphones.

We also present the results for the CHiME-3 dataset. We can see that the margin between using the aligned and unalinged inputs is much smaller in this case. It could be attributed to that most of the

CHiME-3 recording data samples have the speaker almost speaking from the front side with respect to the microphones which results in less variety in directional angles (the common situation when the speaker was asked for holding a tablet to speak the sentence listed in the tablet, and the speech data were recorded through 6 embedded microphones in the same tablet. See the recording setup and recording demonstration picture in CHiME-3 official website) and thus the less spatial diversity, resulting in less benefit of the alignment.

4.3.2 Comparison of AFnet (two-stage) with typical multichannel SE (one-stage)

Having the above insight, we move on to studying the proposed AFnet for exploitation of spatial diversity on the primary dataset with different model and training settings and comparing with typical one-stage deep learning framework approaches. For detailed analysis, we consider four cases below:

i) **U-Net for filter weights:** using only a single U-Net (same model as the Filter Net of AFnet) to estimate the filter weights W_i in typical (one-stage) multichannel SE
ii) **W-Net for filter weights:** using a W-Net (same model as the AFnet, but the output of the first U-Net block goes directly to the next U-Net block without multiplying the input signals) to estimate the filter weights W_i in typical (one-stage) multichannel SE
iii) **AFnet w/o RTF loss:** the two-stage AFnet model trained with normal training scheme without incorporating RTFs
iv) **AFnet:** the two-stage AFnet model trained with the RTF-aware training scheme

Cases i) and ii) represent the typical one-stage approach which uses a DNN to estimate the filter weights for beamforming purposes, serving as baseline methods here. Cases iii) and i) are considered here to observe the role of RTF alignment in the proposed two-stage AFnet.

Figure 2 presents the enhancement results. Comparing case ii) to case i), the W-Net only slightly improves over the U-Net even though having its model size doubled. Case iii), the proposed AFnet model, achieves improvement over the baseline W-Net and U-Net methods for the 8-mic setting even without incorporating RTF information, although in fewer microphone settings it does not perform that well. Finally, the AFnet trained with RTF-aware scheme consistently achieves the best results over all the other cases, suggesting the advantage of incorporating RTF supervision for signal alignment along with the two-stage design.

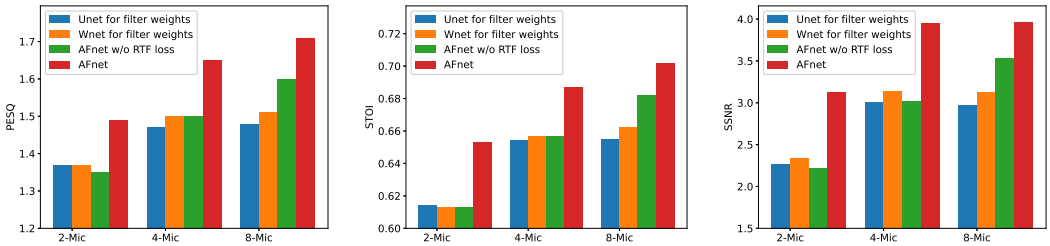

Figure 2: Enhancement performance of different network and training settings. The proposed AFnet performs significantly better than the baseline methods (typical one-stage SE systems) in all cases due to two-stage design with RTF awareness.

4.4 Visualizing the Alignment Outcomes

To further observe the effect of RTF supervision, we visualize the distributions of the alignment masks in Figure 3 from AFnet trained with RTF loss and that without RTF loss. To obtain these plots, we first generate noisy signal samples in which the speech source in each sample is located at one of three pre-defined positions in the room. We feed the noisy samples to the trained AFnet models (both with and without RTF supervision) and obtain the alignment masks for each sample. The distribution of the principal components (PCs) of the obtained masks are plotted. It can be seen that with RTF-aware training, the distributions show clear clusters of speech sources corresponding to the three different positions than with normal training. In addition, with increased number of microphones the clusters also become more separate. The results together with improved enhanced

speech quality provide evidence to the fact that the system is actually exploiting spatial separability of the audio sources for achieving better enhancement performance.

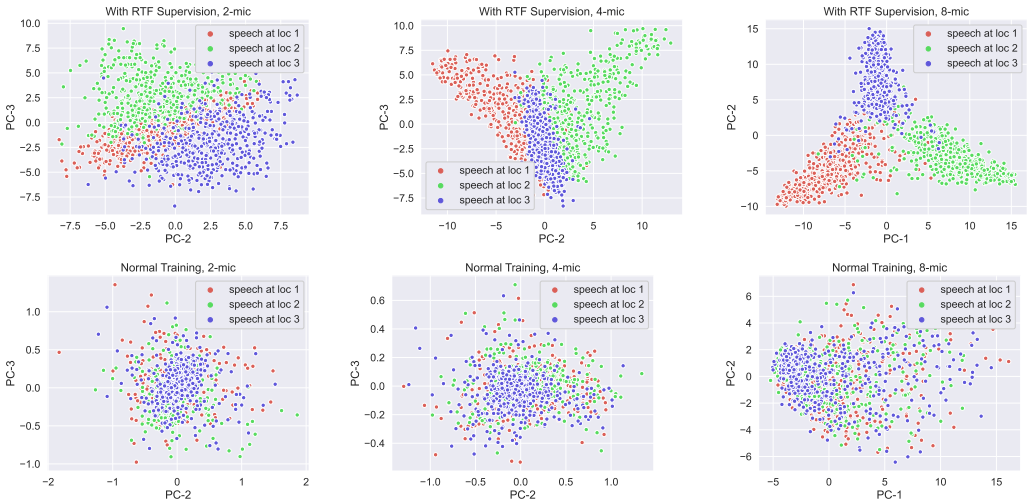

Figure 3: Visualization of the distributions of alignment masks. The RTF-aware training cases (Upper Row) show obvious clusters of signals coming from the three different locations (loc 1, loc 2, loc 3), while the normal training cases (Lower Row) do not present clear boundaries. The results show that incorporating RTFs during training helps the network to better extract the target signal by exploiting spatial separability for signal alignment.

Another angle to inspect the spatial diversity of the dataset can also be provided by observing the alignment masks of the data points. In Figure 4 we compare the distributions of the intermediate alignment masks learned by the AFnet for the primary and CHiME-3 datasets. Here, we plot the distributions of the alignment masks estimated by the trained models for all the test data samples of the two datasets. It can be seen that for the primary, the RTF-aware training leads to a more widespread distribution than that of the normal training case, corresponding to the greater spatial diversity of the target speech source positions. In contrast, in the CHiME-3 dataset the RTF-aware training actually results in a more aggregated cluster than the normal training scheme, indicating that the target speech sources are coming from similar directions. This visualization of the alignment masks provides another angle to inspect the data properties from the spatial diversity aspect for exploring SE performance.

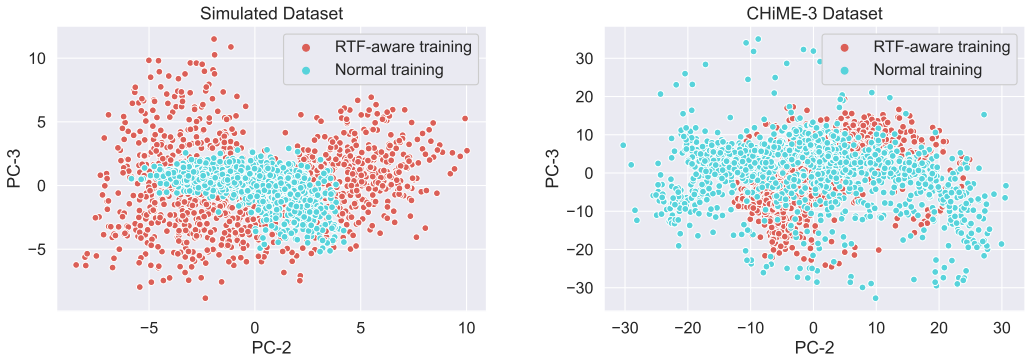

Figure 4: Visualization of the distributions of alignment masks. (Left): the estimated masks for the simulated dataset using AVSpeech audio with Pyroomacoustics. (Right): the estimated masks for the CHiME-3 dataset. Comparing the two cases the results implies that the CHiME-3 dataset has less positional variety of the target source locations than the simulated dataset.

4.4.1 COMPARISON TO SOTA METHODS

To demonstrate the superiority of our method, we compare the AFnet with several SOTA architectures, including: **Conv-TasNet** Luo & Mesgarani (2019): a single-channel SE system performing masking based separation method in the learned transform domain via learnable encoder and decoder. **DCUnet** Choi et al. (2018): a phase-aware single-channel SE approach utilizing complex-valued operations with the U-Net architecture. **FaSNet** Luo et al. (2020): a time-domain, multichannel SE system exploiting cross-channel correlation based on filter-and-sum method.

We train the SOTA models by ourselves.[1] For fairness, all models are trained by minimizing the reconstruction loss $\mathcal{L}_{rec}$ of the power-law compressed MSE, using 512-point FFT with a hop size of 256 via the Hann window. The results are presented in Table 2. It can be seen that the single-channel methods do not perform well even with a larger model size. The FaSNet, with a lighter model, performs slightly better than the two signal channel approaches when the number of microphones is larger. Finally, our proposed AFnet outperforms all the SOTA approaches by significant amount in all the metrics used, while with the smallest model size. The results demonstrate the superiority of the AFnet for enhancing speech in noisy environments.

Table 2: Comparison with SOTA methods. The proposed AFnet outperforms the other existing approaches by a significant amount with the fewest number of parameters.

Methods	# Params	PESQ	STOI	SSNR
Noisy	-	1.23	0.575	-1.54
Conv-TasNet	8.7M	1.28	0.581	0.87
DCUnet	7.6M	1.32	0.615	1.30
FaSNet (2-mic)		1.27	0.579	0.44
FaSNet (4-mic)	2.8M	1.33	0.623	1.08
FaSNet (8-mic)		1.37	0.629	1.37
AFnet (2-mic)		**1.51**	**0.656**	**3.30**
AFnet (4-mic)	**2.6M**	**1.65**	**0.689**	**3.99**
AFnet (8-mic)		**1.69**	**0.698**	**3.93**

Although not an ideal dataset for our study, we still benchmark the AFnet with SOTAs on the CHiME-3 dataset. In Table 3 we report the PESQ and STOI taken from the corresponding papers and the missing entries in the table indicate that the metric is not reported in the reference paper (we do not report SSNR as most papers do not report it). We present both results of the AFnet trained with and without RTF supervision (with the 5-th channel selected as the reference microphone). One can see that the two-stage design of the AFnet is efficient in that it achieves superior performance over other approach while maintaining a smaller model size. The RTF information helps improving the performance, while the gains are minor in this case due to the less spatial diversity of the data. Nevertheless, the results still show the benefit of incorporating signal alignment into the network design, even though on a dataset with limited positional variety of the speech.

Table 3: Comparison with SOTA methods on CHiME-3 dataset.

Methods	# Params	PESQ	STOI
Noisy	-	1.27	0.870
Neural BF Erdogan et al. (2016)	-	2.29	-
MVDR$_{GC}$ Bu et al. (2019)	0.5M	-	0.952
rSDFCN Liu et al. (2020)	2.1M	2.15	0.937
CA Dense U-Net Tolooshams et al. (2020)	>20M	2.44	-
AFnet w/o RTF Loss (ours)	2.6M	2.68	**0.972**
AFnet (ours)		**2.72**	**0.972**

5 CONCLUSION

In this paper, we proposed a novel deep learning based multichannel SE approach utilizing an Align-and-Filter framework that leverages spatial information inherent in the RTFs, which is an essential factor in many conventional signal processing SE algorithms. We found that incorporating signal alignment in the SE model could offer significant improvements, especially when the target speech may come from arbitrary locations. Our findings suggest that signal alignment also plays an important role in deep learning based SE approaches, and that it would be beneficial to leverage such aspect when developing future multichannel SE systems. Though the discussion has been limited to the scenario with only one desired speech source in a noisy environment, extensions to separation of multiple speech sources may be possible and is left for future work.

[1]We use the implementations of Conv-TasNet from: https://github.com/kaituoxu/Conv-TasNet, Large-DCUnet-20 from: https://github.com/mhlevgen/DCUNetTorchSound, and FaSNet-TAC from: https://github.com/yluo42/TAC.

REFERENCES

J. Barker, R. Marxer, E. Vincent, and S. Watanabe. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 504–511, 2015.

S. Braun and I. Tashev. A consolidated view of loss functions for supervised deep learning-based speech enhancement. In *Proceedings of International Conference on Telecommunications and Signal Processing (TSP)*, pp. 72–76, 2021.

S. Bu, Y. Zhao, and M.-Y. Hwang. A novel method to correct steering vectors in MVDR beamformer for noise robust ASR. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 4280–4284, 2019.

J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57 (8):1408–1418, 1969.

H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee. Phase-aware speech enhancement with deep complex U-Net. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2018.

I. Cohen. Relative transfer function identification using speech signals. *IEEE Transactions on Speech and Audio Processing*, 12(5):451–459, 2004.

S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm. Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones. *IEEE Signal Processing Magazine*, 32(2):18–30, 2015.

A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, 37(4):109:1–109:11, 2018.

H. Erdogan, J. R. Hershey, S. Watanabe, M. Mandel, and J. Le Roux. Improved MVDR beamforming using single-channel mask prediction networks. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 1981–1985, 2016.

O. L. Frost. An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, 60(8):926–935, 1972.

S. Gannot, D. Burshtein, and E. Weinstein. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing*, 49(8):1614–1626, 2001.

L. Griffiths and C. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30(1):27–34, 1982.

E. Hadad, F. Heese, P. Vary, and S. Gannot. Multichannel audio database in various acoustic environments. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 313–317, 2014.

R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, and M. Souden. Speech processing for digital home assistants: Combining signal processing with deep-learning techniques. *IEEE Signal Processing Magazine*, 36(6):111–124, 2019.

J. H. L. Hansen and B. L. Pellom. An effective quality evaluation protocol for speech enhancement algorithms. In *Proceedings of International Conference on Speech and Language Processing (ICSLP)*, volume 7, pp. 2819–2822, 1998.

ITU-T Recommendation P.862.2. Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs. *International Telecommunication Union*, 2005.

Z. Koldovskỳ, J. Málek, and S. Gannot. Spatial source subtraction based on incomplete measurements of relative transfer function. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(8):1335–1347, 2015.

A. Krueger, E. Warsitz, and R. Haeb-Umbach. Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):206–219, 2010.

X. Li, L. Girin, R. Horaud, and S. Gannot. Estimation of the direct-path relative transfer function for supervised sound-source localization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2171–2186, 2016.

C.-L. Liu, S.-W. Fu, Y.-J. Li, J.-W. Huang, H.-M. Wang, and Y. Tsao. Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1888–1900, 2020.

P. C. Loizou. *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2007.

Y. Luo and N. Mesgarani. Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8): 1256–1266, 2019.

Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka. End-to-end microphone permutation and number invariant multi-channel speech separation. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6394–6398, 2020.

M. Parchami, W.-P. Zhu, B. Champagne, and E. Plourde. Recent developments in speech enhancement in the short-time Fourier transform domain. *IEEE Circuits and Systems Magazine*, 16(3): 45–77, 2016.

O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pp. 234–241. Springer, 2015.

R. Scheibler, E. Bezzam, and I. Dokmanić. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 351–355, 2018.

C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.

B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy. Channel-attention dense U-Net for multichannel speech enhancement. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 836–840, 2020.

H. L. Van Trees. *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. Honoken, NJ, USA: John Wiley & Sons, 2004.

D. Wang and J. Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018.

Z.-Q. Wang and D. Wang. Mask weighted STFT ratios for relative transfer function estimation and its application to robust ASR. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5619–5623. IEEE, 2018.

K. Wilson, M. Chinen, J. Thorpe, B. Patton, J. Hershey, R. A. Saurous, J. Skoglund, and R. F. Lyon. Exploring tradeoffs in models for low-latency speech enhancement. In *Proceedings of International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 366–370, 2018.

X. Xia and B. Kulis. W-Net: A deep model for fully unsupervised image segmentation. *arXiv preprint arXiv:1711.08506*, 2017.