Interspeech 2018
2-6 September 2018, Hyderabad

All-Neural Multi-Channel Speech Enhancement

Zhong-Qiu Wang[1], *DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA

{wangzhon,dwang}@cse.ohio-state.edu

Abstract

This study proposes a novel all-neural approach for multi-channel speech enhancement, where robust speaker localization, acoustic beamforming, post-filtering and spatial filtering are all done using deep learning based time-frequency (T-F) masking. Our system first performs monaural speech enhancement on each microphone signal to obtain the estimated ideal ratio masks for beamforming and robust time delay of arrival (TDOA) estimation. Then with the estimated TDOA, directional features indicating whether each T-F unit is dominated by the signal coming from the estimated target direction are computed. Next, the directional features are combined with the spectral features extracted from the beamformed signal to achieve further enhancement. Experiments on a two-microphone setup in reverberant environments with strong diffuse babble noise demonstrate the effectiveness of the proposed approach for multi-channel speech enhancement.

Index Terms: beamforming, robust TDOA estimation, spatial filtering, time-frequency masking, deep learning.

1. Introduction

Modern electronic devices typically contain multiple microphones for speech enhancement and robust ASR. With multiple microphones, spatial information can be exploited to complement spectral information for better de-noising and de-reverberation. In spite of decades of efforts, multi-channel speech enhancement remains a major challenge in speech processing. Classical methods are mainly focused on using beamforming to combine multiple signals, and post-filtering for further noise reduction. The beamforming approach designs a linear filter per frequency to boost or maintain the signal from the target direction, while attenuate the interferences from other directions [1]. It typically requires accurate direction of arrival (DOA) and speech or noise covariance matrices estimation. However, commonly used DOA estimation algorithms, such as the generalized cross correlation with phase transform (GCC-PHAT) [2] or the multiple signal classification (MUSIC) [3] algorithms, are not robust enough to environmental noise and room reverberation, as they are only designed to localize the loudest sources in an environment, which may not be the target speaker at all. In environments with strong reverberation and directional or diffuse noise, the summation of the GCC-PHAT coefficients would exhibit high peaks from interference sources or reverberations, and the noise subspace constructed in the MUSIC algorithm would likely not be the true noise subspace. Besides this problem, the microphone geometry is required by the DOA algorithms to derive steering vectors for beamforming. The noise covariance matrices are commonly estimated using leading or ending frames, or silence frames predicted from a voice activity detector. However, conventional voice activity detection algorithms assume that the environmental noise is stationary [4], [5], which is an unrealistic assumption as real-world noises are typically highly non-stationary. Besides these technical difficulties, the noise reduction capability of beamforming is fundamentally limited especially when the number of microphones is restricted and when diffuse noise or room reverberation is present. In addition, it cannot be applied when the sources are spatially close to each other. Conventional post-filtering techniques, which are mainly based on signal statistics and conventional single-channel speech enhancement [5], [1] or spatial filters computed using phase information [6], [7], [8], [1], usually cannot achieve high-quality noise reduction in reverberant multi-source environments.

Recently, deep learning based time-frequency masking has substantially advanced single-channel speech separation and enhancement [9]. The key idea is to train a deep neural network (DNN) to estimate the ideal binary mask (IBM) [10] or the ideal ratio mask (IRM) [11], [12] for enhancement. It has been suggested that the resulting separated speech exhibits remarkable speech intelligibility and quality improvements over conventional methods [13], [14]. To leverage the representational power of deep learning for multi-channel speech enhancement, recent studies encode spatial information as input features for DNN training. In [15], interaural time or level differences (ITD/ILD), and entire cross-correlation coefficients are utilized as extra features to estimate the sub-band IBM in the cochleagram domain. A subsequent study [16] combines ITD, ILD, and the spectral features extracted from a fixed beamformer for further enhancement. A similar study by [17] uses ILD and interchannel phase difference as features to train a deep auto-encoder for enhancement. However, these algorithms assume that the target speech is from a particular direction, typically right in the front in the binaural setup, and therefore may not work well when the target speech is from other directions.

To separate the target speech that could originate from any directions, we first perform robust speaker localization to determine the target direction, and then compute directional features [8], which indicate whether the signal at each T-F unit is from that direction, for DNN training. This way, the DNNs can learn to perform spatial filtering based on the directional features. However, only using directional features is not sufficient enough, as noise and reverberation could also come from the estimated target direction. Therefore, spectral features are also necessary for DNN training so that only the signals from a specific direction and with specific spectral characteristics are enhanced while suppressed otherwise. Clearly, the key step here is the accurate localization of the target speaker. We leverage recent development on T-F masking and deep learning based beamforming [18], [19], [20] for speaker localiza-

This research was supported in part by an AFRL contract (FA8750-15-1-0279), an NSF grant (IIS-1409431), and the Ohio Supercomputer Center.

10.21437/Interspeech.2018-1664

tion. The proposed localization and enhancement algorithms exhibit strong robustness in our experiments.

The rest of this paper is organized as follows. We describe the proposed algorithm in Section 2. Experimental setup and evaluation results are presented in Section 3 and 4. Section 5 concludes this paper.

2. System Description

We first introduce the beamforming algorithms based on deep learning and then present our algorithm for TDOA estimation. The estimated time delay is used to compute directional features, which are then combined with spectral features for further enhancement. See Figure 1 for an overall illustration.

2.1. MVDR Beamforming Based on T-F Masking

Suppose that there is only one target speaker, the physical model for a pair of signals received by a two-microphone array in noisy and reverberant environments is assumed to have the following form:

$$\boldsymbol{y}(t,f) = \boldsymbol{c}(f)s(t,f) + \boldsymbol{h}(t,f) + \boldsymbol{n}(t,f) \tag{1}$$

where $s(t,f)$ is the STFT value of the target source signal at time t and frequency f, $\boldsymbol{c}(f)$ is the acoustic transfer function from the sound source to the array, $\boldsymbol{c}(f)s(t,f)$ and $\boldsymbol{h}(t,f)$ are the direct sound and early and late reverberation of the target signal, and $\boldsymbol{y}(t,f)$ and $\boldsymbol{n}(t,f)$ represent the received mixture signal and the received reverberant noise component.

Recent studies in the CHiME challenges [21], [22] suggest that the speech and noise statistics critical for accurate beamforming can be well-estimated using deep learning based T-F masking [19], [18], [23]. The key advance is to use a powerful DNN to identify speech-dominated and noise-dominated T-F units so that the speech covariance matrices can be computed from speech-dominated T-F units and the noise covariance matrices from noise-dominated T-F units. Remarkable improvements in terms of ASR performance have been observed over conventional beamforming approaches [21], [22], [24].

Following [19], [20], we estimate the speech and noise covariance matrices as follows:

$$\widehat{\Phi}_s(f) = \frac{1}{\sum_t \eta(t,f)} \sum_t \eta(t,f)\boldsymbol{y}(t,f)\,\boldsymbol{y}(t,f)^H \tag{2}$$

$$\widehat{\Phi}_n(f) = \frac{1}{\sum_t \xi(t,f)} \sum_t \xi(t,f)\boldsymbol{y}(t,f)\boldsymbol{y}(t,f)^H \tag{3}$$

where $(\cdot)^H$ represents conjugate transpose, and $\eta(t,f)$ and $\xi(t,f)$ are the weighting terms denoting the importance of each T-F unit for the speech and noise covariance matrices computation. They are defined as the product of individual estimated T-F masks:

$$\eta(t,f) = \prod_{i=1}^{D} \widehat{M}_i(t,f) \tag{4}$$

$$\xi(t,f) = \prod_{i=1}^{D} \left(1 - \widehat{M}_i(t,f)\right) \tag{5}$$

where D(=2 in this study) is the number of microphones and $\widehat{M}_i(t,f)$ is the estimated mask of microphone signal i.

Assuming that the first microphone is the reference microphone, the relative transfer function is estimated as:

$$\hat{\boldsymbol{c}}(f) = \mathcal{P}\{\widehat{\Phi}_s(f)\} = [\frac{1}{\sqrt{\hat{A}(f)^2+1}}, \frac{\hat{A}(f)}{\sqrt{\hat{A}(f)^2+1}} e^{j\hat{\theta}(f)}]^T \tag{6}$$

where $\mathcal{P}\{\cdot\}$ computes the principal eigenvector, and $\hat{A}(f)$ and $\hat{\theta}(f)$ are the estimated level and phase difference, respectively. The rationale is that if $\widehat{\Phi}_s(f)$ is well-estimated, it would be close to a symmetric rank-one matrix [1] as the target speech is from a directional speaker source. In such a case, the principal eigenvector is a reasonably good estimate of the relative transfer function [18], [20], [25].

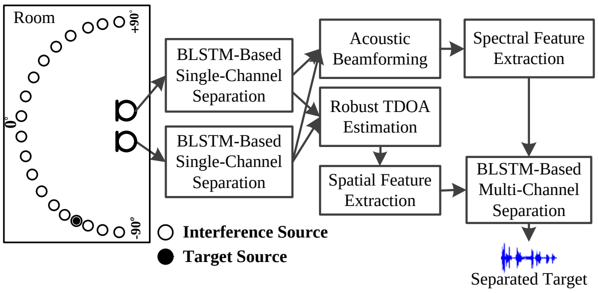
Room
+90°
0°
-90°
BLSTM-Based Single-Channel Separation
BLSTM-Based Single-Channel Separation
Acoustic Beamforming
Robust TDOA Estimation
Spatial Feature Extraction
Spectral Feature Extraction
BLSTM-Based Multi-Channel Separation
○ Interference Source
● Target Source
Separated Target

Figure 1. Illustration of overall system.

With $\widehat{\Phi}_n(f)$ and $\hat{\boldsymbol{c}}(f)$ estimated, a minimum variance distortion-less response (MVDR) beamformer is constructed:

$$\widehat{\boldsymbol{w}}(f) = \frac{\widehat{\Phi}_n(f)^{-1}\hat{\boldsymbol{c}}(f)}{\hat{\boldsymbol{c}}(f)^H\widehat{\Phi}_n(f)^{-1}\hat{\boldsymbol{c}}(f)} \tag{7}$$

and enhancement results are obtained using:

$$\hat{y}_{bf}(t,f) = \widehat{\boldsymbol{w}}(f)^H\boldsymbol{y}(t,f) \tag{8}$$

Since beamforming algorithms only perform linear filtering per frequency, it is typically incapable of achieving sufficient enhancement, especially when the target source is spatially close to noise sources and diffuse noise or room reverberation is present. To improve the performance, our study uses deep learning based T-F masking as a post-filter for further enhancement. We extract spectral features from the beamformed signal, $\hat{y}_{bf}(t,f)$, to estimate another T-F mask. This mask is then element-wisely multiplied with $\hat{y}_{bf}(t,f)$ to get the enhancement result.

2.2. Robust TDOA Estimation

The estimated steering vectors contain sufficient information about the interchannel level and phase differences of the target speech at each frequency [1], [26]. This section seeks a way to extract the time delay information from the estimated steering vectors. As they are computed independently at each frequency using eigendecomposition, the estimated phase delay, $\hat{\theta}(f)$, would not strictly follow a linear phase structure. In our study, we propose to enumerate a set of potential time delays of interests and find a time delay that maximizes the following objective function.

$$Similarity(\tau) = \sum_f cos\left(\hat{\theta}(f) - (-2\pi\frac{f}{N}f_s\tau)\right) \tag{9}$$

$$\hat{\tau} = argmax_\tau\, Similarity(\tau) \tag{10}$$

where N is the number of DFT frequencies, f_s is the sampling rate, and τ is a hypothesized time delay in seconds. Note that f ranges from 0 to $N/2$. Intuitively, this algorithm searches for a time delay τ with its hypothesized phase difference, $-2\pi\frac{f}{N}f_s\tau$, best matched with $\hat{\theta}(f)$ at all the frequency bands.

An alternative objective function is to put more weights on the frequencies with higher SNR, as some frequencies may be particularly bad due to the specific characteristics of environmental noise and room reverberation.

$$\gamma(f) = \sum_t \eta(t,f) \tag{11}$$

$$Similarity(\tau) = \sum_f \gamma(f)cos\left(\hat{\theta}(f) - (-2\pi\frac{f}{N}f_s\tau)\right) \tag{12}$$

where $\eta(t,f)$ is defined as in Eq. (4). The rationale of using Eq. (11) is that if the estimated mask values are high, it is likely that the SNR is also high.

There are previous attempts [27], [28], [29] at deriving time delays from estimated steering vectors at each frequency or each T-F unit. They directly divide the phase difference at each T-F by its angular frequency. However, doing it this way ignores the fact that there could be multiple time delays giving exactly the same phase difference due to phase wrapping and spatial aliasing. The proposed method addresses this ambiguity by checking all the time delays of interests and using their similarity scores to the phase delay of the estimated steering vectors to determine the underlying time delay.

2.3. Directional Features

After obtaining the estimated time delay, $\hat{\tau}$, we compute the spatial features following [8]:

$$DF(t,f) = \cos\left(\angle y_1(t,f) - \angle y_2(t,f) - 2\pi \frac{f}{N} f_s \hat{\tau}\right), \quad (13)$$

where subscript 1 and 2 index microphone channels. The rationale is that if $\hat{\tau}$ is accurate enough, $DF(t,f)$ would be close to one if the T-F unit is dominated by the signal coming from the estimated target direction, and much less than one otherwise. It can therefore be used as input features for DNN training to enhance the signals coming from the estimated direction and filter out signals, typically noise and reverberation, from other directions. In addition, using this feature alone is not sufficient enough, as noise or reverberation could also come from the estimated direction. To address this issue, spectral features are still indispensable. Our system hence integrates both the spatial features and spectral features for DNN training such that only the signals coming approximately from a specific direction and with specific spectral characteristics are enhanced or maintained, while filtered out otherwise. This approach distinguishes our study with [8], which only uses spatial information for enhancement and does not address speaker localization in a robust way.

The spectral features can be extracted from the received mixture signal at each microphone as well as from the beamformed signal obtained using Eq. (8).

2.4. Mask Estimation

The accuracy of mask estimation is essential in the proposed framework. We use the direct sound signal as the target and the rest as the noise to define the IRM:

$$\text{IRM}_i(t,f) = \frac{|c_i(f)s(t,f)|^2}{|c_i(f)s(t,f)|^2 + |h_i(t,f) + n_i(t,f)|^2} \quad (14)$$

Recent studies suggest that bi-directional long short-term memory (BLSTM) usually leads to consistently better mask estimation results over many other neural networks [30]. In our study, two BLSTMs are trained for mask estimation, one only taking in single-channel spectral information and the other one taking in both spectral and spatial features, as is illustrated in Figure 1. The estimated masks are applied to the unprocessed signal or the beamformed signal for enhancement.

3. Experimental Setup

The proposed algorithms are evaluated using a two-microphone setup for speech enhancement in highly reverberant environments with strong diffuse babble noise. An illustration of the experimental setups is shown in Figure 1. A room impulse response (RIR) generator[1] based on the image method [31] is employed to generate the RIRs. For the training and validation data, we put one interference speaker at each of the 36 directions ranging from −87.5° to 87.5° in steps of 5°, and the target speaker at one of the 36 directions. For the testing data, we put one inference speaker at each of the 37 directions spanning from −90° to 90° in steps of 5°, and the target speaker at one of the directions. This way, the testing RIRs are unseen during training. The distance between each speaker to the array center is 1.0m. The room size is at 8x8x3m, and the two microphones are placed around the center of the room. The distance between the two microphones is 0.2m and the heights are both set to 1.5m. The T60 of each mixture is randomly picked from 0.0s to 1.0s in steps of 0.1s. The 720 IEEE female utterances [32] are utilized as the target in our experiments. We randomly split them into 500, 100 and 120 utterances to generate the training, validation and testing data. To create the diffuse babble noise, we first concatenate the utterances of each of the 630 speakers in the TIMIT dataset, and then randomly pick 37 (or 36) speech segments from 37 (or 36) randomly-chosen speakers to put at each of the 37 (or 36) directions. For each speaker in the babble noise, we use the first half of the concatenated utterance to generate the training and validation babble noise and the second half to generate the testing babble noise. There are 25,000, 800, and 3,000 two-channel mixtures in the training, validation and testing set, respectively. The average duration of the mixtures is 2.4s. The input SNR computed from reverberant speech and reverberant noise is fixed at -6dB. Note that if the direct sound signal is considered as the target speech in SNR computation, the SNR is even lower, depending on the direct-to-reverberant energy ratio (DRR).

We train our BLSTMs using all the single-channel signals together with their oracle beamformed signals. We use the IRM to compute the weights in Eq. (4) and (5) to derive the oracle beamformed signals. For each two-channel mixture, we can use the first microphone signal as well as the second microphone signal as the reference for beamforming, so there are two beamformed signals created for training. For each beamformed signal, we use the beamformed speech together with the beamformed noise to define its IRM. The BLSTM is trained using 100,000 (=25,000*2 + 25,000*2) mixtures in total. The BLSTM for single-channel enhancement is trained using log power spectrogram features, and the BLSTM for multi-channel enhancement is trained using the concatenation of log power spectrogram features and directional features. Similarly, we use the IRM to derive the oracle directional features for model training, while at run time, estimated IRMs are utilized for beamforming and directional feature computation.

Both BLSTMs consist of two hidden layers each with 384 units in each direction. Adam is used for optimization. The window size is 32ms and the hop size is 8ms. The sampling rate is 16 kHz. After hamming window is applied, 512-point FFT is performed to extract 257-dimensional log power spectrogram feature of each frame for BLSTM training. The input dimension of the BLSTM for single-channel enhancement is therefore 257, while 514 (=257*2) for the other BLSTM. Sigmoidal activations are used in the output layer. Sentence-level mean normalization is performed on the spectral features before global mean-variance normalization to alleviate the effects of reverberation. Only global mean-variance normalization is performed on the directional features.

We measure speaker localization performance in terms of gross accuracy, which considers a prediction to be correct if the prediction is within 5^o (inclusive) from the true target direction. At run time, we only perform enhancement on the first channel signal and use the direct sound signal at the first channel as the reference for metric computation. We evaluate

[1] See https://github.com/ehabets/RIR-Generator.

Table 1. Comparison of TDOA estimation performance (% Gross Accuracy) of different approaches in the two-microphone setup.

Approaches	AVG	T60(s)/DRR(dB)									
		0.0/-	0.2/7.2	0.3/3.0	0.4/0.9	0.5/-0.5	0.6/-1.6	0.7/-2.5	0.8/-3.2	0.9/-3.9	1.0/-4.4
GCC-PHAT	25.8	40.4	39.9	33.9	37.4	25.2	19.4	20.1	15.8	13.4	13.4
TDOA Estimation from Steering Vectors (using Eq. (9))	89.0	94.9	94.1	96.3	91.7	92.2	90.3	84.6	82.6	81.3	82.2
TDOA Estimation from Steering Vectors (using Eq. (12))	**92.0**	96.5	96.5	97.3	95.5	94.6	93.1	89.0	88.4	86.3	83.3
Using IRM to Get Oracle $\hat{\boldsymbol{c}}(f)$ (using Eq. (12))	99.8	100.0	100.0	100.0	100.0	100.0	100.0	99.7	100.0	98.7	100.0

Table 2. Comparison of STOI (%) results of different approaches in the two-microphone setup.

Approaches	Estimated Mask Applied on ?	AVG	T60(s)/DRR(dB)									
			0.0/-	0.2/7.2	0.3/3.0	0.4/0.9	0.5/-0.5	0.6/-1.6	0.7/-2.5	0.8/-3.2	0.9/-3.9	1.0/-4.4
Unprocessed	-	48.5	54.0	51.9	50.3	49.3	48.6	47.4	46.5	46.0	45.5	44.9
Single-Channel BLSTM (logps from $y_1(t,f)$)	$y_1(t,f)$	67.4	76.0	73.4	70.9	69.5	67.9	66.2	64.5	63.0	61.7	60.8
Multi-Channel BLSTM (logps from $y_1(t,f)$ + DF)	$y_1(t,f)$	71.4	79.6	77.7	75.3	74.1	72.1	70.7	68.5	66.5	65.6	64.1
Multi-Channel BLSTM (logps from $y_1(t,f)$ + oracle DF)	$y_1(t,f)$	71.8	79.8	77.9	75.4	74.3	72.3	70.9	69.1	67.3	66.3	65.1
T-F Masking Based Beamforming	-	54.3	61.5	59.3	57.2	55.6	54.6	52.9	51.8	50.8	50.0	49.2
Oracle Beamforming (oracle $\widehat{\Phi}_s(f)$ and $\widehat{\Phi}_n(f)$)	-	55.6	62.1	60.1	58.3	56.8	56.0	54.2	53.4	52.5	51.6	50.9
Single-Channel BLSTM (logps from $\hat{y}_{bf}(t,f)$)	$\hat{y}_{bf}(t,f)$	73.1	81.9	79.8	77.0	75.2	73.9	71.8	70.0	68.3	67.2	66.1
Multi-Channel BLSTM (logps from $\hat{y}_{bf}(t,f)$ + DF)	$\hat{y}_{bf}(t,f)$	74.4	82.7	81.0	78.6	76.9	75.0	73.5	71.5	69.2	68.2	67.1
Multi-Channel BLSTM (logps from $\hat{y}_{bf}(t,f)$ + oracle DF)	$\hat{y}_{bf}(t,f)$	74.8	82.9	81.1	78.7	77.1	75.2	73.8	71.8	69.9	69.0	68.2

Table 3. Comparison of PESQ results of different approaches in the two-microphone setup.

Approaches	Estimated Mask Applied on ?	AVG	T60(s)/DRR(dB)									
			0.0/-	0.2/7.2	0.3/3.0	0.4/0.9	0.5/-0.5	0.6/-1.6	0.7/-2.5	0.8/-3.2	0.9/-3.9	1.0/-4.4
Unprocessed	-	0.98	1.00	1.03	1.01	0.98	0.99	0.95	0.97	0.94	0.96	0.96
Single-Channel BLSTM (logps from $y_1(t,f)$)	$y_1(t,f)$	1.77	2.02	1.97	1.92	1.86	1.79	1.73	1.68	1.61	1.58	1.51
Multi-Channel BLSTM (logps from $y_1(t,f)$ + DF)	$y_1(t,f)$	1.91	2.17	2.15	2.07	2.02	1.93	1.87	1.81	1.74	1.69	1.63
Multi-Channel BLSTM (logps from $y_1(t,f)$ + oracle DF)	$y_1(t,f)$	1.92	2.18	2.15	2.07	2.03	1.94	1.88	1.83	1.76	1.71	1.65
T-F Masking Based Beamforming	-	1.08	1.20	1.19	1.13	1.10	1.09	1.04	1.04	1.03	1.00	0.97
Oracle Beamforming (oracle $\widehat{\Phi}_s(f)$ and $\widehat{\Phi}_n(f)$)	-	1.08	1.19	1.20	1.16	1.13	1.09	1.04	1.04	1.01	0.98	0.99
Single-Channel BLSTM (logps from $\hat{y}_{bf}(t,f)$)	$\hat{y}_{bf}(t,f)$	2.01	2.27	2.25	2.18	2.10	2.03	1.96	1.91	1.84	1.79	1.73
Multi-Channel BLSTM (logps from $\hat{y}_{bf}(t,f)$ + DF)	$\hat{y}_{bf}(t,f)$	2.05	2.31	2.31	2.23	2.17	2.08	2.02	1.95	1.88	1.81	1.79
Multi-Channel BLSTM (logps from $\hat{y}_{bf}(t,f)$ + oracle DF)	$\hat{y}_{bf}(t,f)$	2.06	2.32	2.31	2.23	2.17	2.08	2.02	1.96	1.89	1.83	1.80

the enhancement performance using the short-time objective intelligibility (STOI) and perceptual estimation of speech quality (PESQ) measures, which are the objective measures of speech intelligibility and quality, respectively.

4. Evaluation Results

Since the accuracy of TDOA estimation is critical for the quality of the directional features, we first report the performance of the proposed algorithm for TDOA estimation in each reverberation level together with the DRR in Table 1. The results obtained using oracle information is marked in grey. The conventional GCC-PHAT algorithm [33] is used as the baseline for comparison. Since the target speaker is fixed within each utterance, we sum the GCC coefficients over all the T-F units to get the estimated time delay. Its performance, however, is only 25.8% gross accuracy in our experimental setup. The proposed TDOA estimation algorithm substantially improves the performance to 92.0% gross accuracy. In addition, the weighting mechanism as in Eq. (12) also leads to significant improvement from 89.0% to 92.0% gross accuracy. Interestingly, if the IRM is used to compute $\hat{\boldsymbol{c}}(f)$, almost perfect gross accuracy can be obtained. This indicates the strong potential of the proposed TDOA algorithm.

We then report the STOI and PESQ results in Table 2 and 3. As can be seen from the first two entries, single-channel enhancement achieves large improvements (from 48.5% to 67.4% for STOI and from 0.98 to 1.77 for PESQ), even only using spectral information. Incorporating the directional features for multi-channel enhancement significantly improves the STOI from 67.4% to 71.4% and PESQ from 1.77 to 1.91. The fourth entry provides the performance when using oracle directional features obtained by using true target directions. As can be observed from entry 3 and 4, the performances are similar, likely because the proposed TDOA estimation algorithm can already accurately determine the target direction in our experiments.

Using the T-F masking based beamforming only gets slight improvement in such a challenging environment with strong reverberation and noise (from 48.5% to 54.3% for STOI and from 0.98 to 1.08 for PESQ). Nonetheless, although estimated speech and noise statistics are used, its performance is close to the oracle MVDR beamforming results obtained using oracle covariance matrices, indicating the effectiveness of deep learning based T-F masking for beamforming.

Applying the single-channel BLSTM on top of the beamforming results reaches 73.1% STOI and 2.01 PESQ from 54.3% and 1.08. Further adding spatial features yields slight improvement. As can be seen from the results, including a beamforming module by extracting spectral features from beamformed signals, $\hat{y}_{bf}(t,f)$, and applying the estimated mask on $\hat{y}_{bf}(t,f)$ leads to consistent improvement than using unprocessed $y_1(t,f)$. This is possibly because beamforming algorithms can already suppress the noise and enhance the noisy phase to some extent.

5. Concluding Remarks

This study has proposed a novel framework for multi-channel speech enhancement based on time-frequency masking and deep learning. The key step is to leverage the power of deep learning based T-F masking to accurately compute the statistics for beamforming and estimate the target direction, so that spectral and spatial information can be utilized simultaneously to enhance the signal from a specific direction and with specific spectral characteristics. The proposed framework is flexible and versatile enough to be extended to arrays with more than two microphones. Future research would evaluate the performance of the proposed algorithm on robust ASR tasks. We shall also consider performing de-noising and de-reverberation in a two-stage way as in our recent study [34].

6. References

[1] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A Consolidated Perspective on Multi-Microphone Speech Enhancement and Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 692–730, 2017.

[2] C. Knapp and G. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320–327, 1976.

[3] R. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[4] Y. Hu and P. C. Loizou, "A Comparative Intelligibility Study of Single-Microphone Noise Reduction Algorithms," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1786, 2007.

[5] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC press, 2013.

[6] M. L. Seltzer and I. Tashev, "A Log-MMSE Adaptive Beamformer using a Nonlinear Spatial Filter," in *Proceedings of IWAENC*, 2008.

[7] I. Tashev and A. Acero, "Microphone Array Post-Processor using Instantaneous Direction of Arrival," in *Proceedings of IWAENC*, 2006.

[8] P. Pertilä and J. Nikunen, "Distant Speech Separation using Predicted Time-Frequency Masks from Spatial Features," *Speech Communication*, vol. 68, pp. 97–106, 2015.

[9] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2018.

[10] D. L. Wang and G. J. Brown, *Eds., Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley-IEEE Press, 2006.

[11] A. Narayanan and D. L. Wang, "Ideal Ratio Mask Estimation using Deep Neural Networks for Robust Speech Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7092–7096.

[12] Z.-Q. Wang and D. L. Wang, "Recurrent Deep Stacking Networks for Supervised Speech Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 71–75.

[13] E. Healy, S. Yoho, Y. Wang, and D. L. Wang, "An Algorithm to Improve Speech Recognition in Noise for Hearing-Impaired Listeners," *The Journal of the Acoustical Society of America*, vol. 23, no. 6, pp. 3029–3038, 2013.

[14] Y. Wang, A. Narayanan, and D. L. Wang, "On Training Targets for Supervised Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[15] Y. Jiang, D. L. Wang, R. Liu, and Z. Feng, "Binaural Classification for Reverberant Speech Segregation using Deep Neural Networks," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 22, no. 12, pp. 2112–2121, 2014.

[16] X. Zhang and D. L. Wang, "Deep Learning Based Binaural Speech Separation in Reverberant Environments," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 25, no. 5, pp. 1075–1084, 2017.

[17] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring Multi-Channel Features for Denoising-Autoencoder-Based Speech Enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 116–120.

[18] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 System: Advances in Speech Enhancement and Recognition for Mobile Multi-Microphone Devices," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015, pp. 436–443.

[19] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM Supported GEV Beamformer Front-End for the 3rd CHiME Challenge," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015, pp. 444–451.

[20] X. Zhang, Z.-Q. Wang, and D. L. Wang, "A Speech Enhancement Algorithm by Iterating Single- and Multi-Microphone Processing and its Application to Robust ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 276–280.

[21] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The Third 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015, pp. 504–511.

[22] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The Third 'CHiME' Speech Separation and Recognition Challenge: Analysis and Outcomes," *Computer Speech and Language*, vol. 46, pp. 605–626, 2017.

[23] Z.-Q. Wang and D. Wang, "Mask Weighted STFT Ratios for Relative Transfer Function Estimation and its Application to Robust ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[24] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller, "Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments," *arXiv preprint arXiv:1705.10874*, May 2017.

[25] Z.-Q. Wang and D. L. Wang, "On Spatial Features for Supervised Speech Separation and its Application to Beamforming and Robust ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[26] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust TDOA Estimation Based on Time-Frequency Masking and Deep Neural Networks," in *Proceedings of Interspeech*, 2018.

[27] S. Rickard and O. Yilmaz, "On the Approximate W-disjoint Orthogonality of Speech," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 1529–1532, 2002.

[28] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA Estimation for Multiple Sparse Sources with Normalized Observation Vector Clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006, pp. 33–36.

[29] N. T. N. Tho, S. Zhao, and D. L. Jones, "Robust DOA Estimation of Multiple Speech Sources," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 2287–2291.

[30] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation*, 2015, pp. 91–99.

[31] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *Journal of Acoustical Society of America*, vol. 65, no. 4, p. 943, 1979.

[32] IEEE, "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[33] J. DiBiase, H. Silverman, and M. Brandstein, "Robust Localization in Reverberant Rooms," in *Microphone Arrays*, Berlin Heidelberg: Springer, 2001, pp. 157–180.

[34] Y. Zhao, Z.-Q. Wang, and D. L. Wang, "A Two-Stage Algorithm for Noisy and Reverberant Speech Enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 5580–5584.