Interspeech 2018
2-6 September 2018, Hyderabad

Robust TDOA Estimation Based on Time-Frequency Masking and Deep Neural Networks

Zhong-Qiu Wang[1], *Xueliang Zhang*[3], *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
[3]Department of Computer Science, Inner Mongolia University, China
{wangzhon,dwang}@cse.ohio-state.edu, cszxl@imu.edu.cn

Abstract

Deep learning based time-frequency (T-F) masking has dramatically advanced monaural speech separation and enhancement. This study investigates its potential for robust time difference of arrival (TDOA) estimation in noisy and reverberant environments. Three novel algorithms are proposed to improve the robustness of conventional cross-correlation-, beamforming- and subspace-based algorithms for speaker localization. The key idea is to leverage the power of deep neural networks (DNN) to accurately identify T-F units that are relatively clean for TDOA estimation. All of the proposed algorithms exhibit strong robustness for TDOA estimation in environments with low input SNR, high reverberation and low direction-to-reverberant energy ratio.

Index Terms: GCC-PHAT, time-frequency masking, robust TDOA estimation, deep neural networks.

1. Introduction

Robust speaker localization has a lot of applications in real-world products, such as human-computer interaction, robotics and beamforming. Conventionally, the generalized cross correlation with phase transform (GCC-PHAT) [1] or steered-response power with phase transform (SRP-PHAT) [2] algorithm and the multiple signal classification (MUSIC) [3] algorithm are the two most popular techniques in sound source localization. However, 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, and the noise subspace constructed from the eigenvectors corresponding to the smallest eigenvalues of noisy speech covariance matrices in the MUSIC algorithm would likely not be the true noise subspace.

To improve the robustness, earlier studies apply SNR weighting to emphasize the frequencies with higher SNR for the GCC-PHAT algorithm. Many SNR estimation algorithms have been applied, such as the rule-based methods [4], voice activity detection based algorithms [5], or minimum mean square error based approaches [6]. However, these algorithms usually assume that the noise is stationary, which is an unrealistic assumption in real-world environments. In the area of computational auditory scene analysis, it has been suggested that human may perform sound localization by first performing speech separation [7]. Motivated by this auditory observation, we approach the speaker localization problem from the angle of monaural speech separation.

Our key observation is that even for a severely corrupted signal, there are still many T-F units dominated by target speech [7]. According to Woodruff and Wang [8], [9], these T-F units, with much cleaner phase, are sufficient enough for robust speaker localization. In other words, this formulation aims at assigning binary labels to T-F units according to speech and noise dominance. A profound consequence of this new formulation is that robust speaker localization can now be approached from the angle of T-F unit level classification (or regression) using modern machine learning techniques. However, previous attempts, which employ Gaussian mixture models and support vector machines [8], [9], [10], [11], are incapable of accurately finding these T-F units in a haystack of noise and reverberation. In recent years, monaural speech separation techniques based on T-F masking and deep learning have demonstrated strong potential and overwhelming advantages over conventional speech separation and enhancement algorithms [12], [13]. It has been validated in many studies that, with the strong learning power of DNNs, this approach is capable of accurately determining the speech or noise dominance at each T-F unit [14], and the resulting separated speech exhibits remarkable speech intelligibility and quality improvements [15].

In this context, we propose three novel algorithms based on T-F masking and deep learning to improve conventional cross-correlation-, beamforming- and subspace-based algorithms for speaker localization. In the literature, there are recent studies applying deep learning based T-F masking for robust speaker localization [16], [17]. We note that our study is independently developed and we not only explore the classic algorithms, such as GCC-PHAT, but also propose new algorithms to utilize the estimated masks from DNNs for better speaker localization in noisy and reverberant environments. Besides the T-F masking based approach, another popular deep learning based method for sound localization is to discretize potential directions and formulate speaker localization as a supervised classification problem [18], [19]. However, such approaches generally suffer from resolution problems and microphone geometry mismatches. In contrast, the masking based approach is more flexible and versatile, as the DNN only needs to determine the speech or noise dominance at each T-F unit, which is a well-defined and well-studied task in monaural speech separation [13].

2. System Description

We first review the GCC-PHAT algorithm in Section 2.1 and then present the proposed algorithms. The discussion on deep learning based mask estimation is in Section 2.5.

2.1. GCC-PHAT

Assuming that there is only one target source, the physical model for a pair of signals in reverberant and noisy environments can be formulated as:

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

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.2018-1652

where $s(t,f)$ represents the STFT value of the target source at time t and frequency f, $\boldsymbol{c}(f)$ is the relative transfer function, and $\boldsymbol{c}(f)s(t,f)$, $\boldsymbol{h}(t,f)$, $\boldsymbol{n}(t,f)$, and $\boldsymbol{y}(t,f)$ are the STFT vectors of the direct sound, early and late reverberation, reverberant noise and received mixture, respectively. By choosing the first microphone as the reference, the relative transfer function $\boldsymbol{c}(f)$ is assumed to have the following form:

$$\boldsymbol{c}(f) = \left[1, A(f)e^{-j2\pi\frac{f}{N}f_s\tau^*}\right]^T, \tag{2}$$

where τ^* is the underlying time delay in seconds, $A(f)$ is a real-valued gain, f_s is the sampling rate in Hz, N is the number of DFT frequencies, and $[\cdot]^T$ stands for matrix transpose. Note that f ranges from 0 to $N/2$.

The GCC-PHAT algorithm [1], [2] estimates the time delay by computing the generalized cross-correlation function with a weighting mechanism based on the phase transform:

$$\begin{aligned} GCC_{PHAT}(t,f,\tau) &= \mathcal{R}eal\left\{\frac{y_1(t,f)y_2(t,f)^H}{|y_1(t,f)||y_2(t,f)^H|}e^{-j2\pi\frac{f}{N}f_s\tau}\right\} \\ &= \cos\left(\angle y_1(t,f) - \angle y_2(t,f) - 2\pi\frac{f}{N}f_s\tau\right), \end{aligned} \tag{3}$$

where $(\cdot)^H$ stands for conjugate transpose, $\mathcal{R}eal\{\cdot\}$ extracts the real part, $|\cdot|$ computes the magnitude, and subscripts 1 and 2 index microphone channel. Intuitively, this algorithm first tries to align these two signals using a candidate time delay τ and then calculates their cosine distance. If the cosine distance is close to one, it means that the candidate time delay is close to the true time delay. Each GCC coefficient is therefore in between -1 and 1. Assuming that the sound source is fixed within each utterance, the GCC coefficients are then pooled together and the τ giving the largest summation is considered as the time delay estimate. We emphasize that the PHAT weighting [20] is essential here. If the normalization is not performed, some frequencies with higher energy would have larger GCC coefficients and dominate the summation.

2.2. Mask-Weighted GCC-PHAT

Although GCC-PHAT performs well in moderately reverberant environments, it collapses even in slightly noisy environments. To improve the robustness, we include a masking-based weighting term into the algorithm following [21], [22]:

$$GCC_{PHAT-MASK}(t,f,\tau) = \eta(t,f)GCC_{PHAT}(t,f,\tau), \tag{4}$$

where $\eta(t,f)$ denotes the importance of the T-F pair for TDOA estimation. It is defined as:

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

where D(=2 in this case) is the number of microphone channels and $\widehat{M}_i$ is the estimated mask representing the estimated speech energy portion at each T-F unit of signal i. The time delay is estimated using:

$$\hat{\tau} = arg\max_{\tau}\sum_{t,f} GCC_{PHAT-MASK}(t,f,\tau) \tag{6}$$

The weighting mechanism will put more weights on the T-F units dominated by target speech across all the microphone channels, as they contain much cleaner phase for localization. Adding this weighting term would hence likely sharpen the peak corresponding to target speech in the summation and suppress the peaks corresponding to noise sources, only if the masks can be accurately estimated.

Note that our study estimates one mask from each single-channel signal and then combine them using the product as the weights. The resulting neural network for mask estimation would be directly applicable to microphone arrays with various numbers of microphones and microphone geometry, while microphone geometry information is still required to estimate the three-dimensional location.

Following [21], [22], a recent study by Pertila *et al.* [16] also proposed to use neural network based T-F masking to improve the SRP-PHAT algorithm. Their system first averages the log magnitude features from all the channels and then estimates a mask from the averaged features using a deep convolutional neural network. Subsequently, the estimated mask is directly used as the weights for the classic SRP-PHAT algorithm. Here, we point out that doing it this way could include unreliable T-F units for localization. Because the input SNRs at different microphone channels are normally different, averaging the log magnitudes would lead to a signal with an SNR in between the SNRs of the original signals. The resulting estimated mask would likely contain estimated speech dominant T-F units that are not speech dominant across all the microphone signals. In contrast, our system estimates a mask from each channel individually, and then combines them using the product. The product operation would automatically identify and only put more weights on T-F units dominated by speech across all the microphone channels.

2.3. Mask-Weighted Steered-Response SNR

The proposed mask-weighted GCC-PHAT algorithm uses a weighting mechanism to emphasize speech dominated T-F units so that the steered-response power of target speech, rather than noisy speech is used as the indicator of target direction. Following [23], this section explores the use of steered-response SNR as the indicator, as SNR considers not only speech power but also noise power.

We first compute the speech covariance matrix, $\widehat{\Phi}_s(f)$, and noise covariance matrix, $\widehat{\Phi}_n(f)$, at each frequency in the following way [24], [25], [26].

$$\widehat{\Phi}_s(f) = \frac{1}{\sum_t \eta(t,f)}\sum_t \eta(t,f)\boldsymbol{y}(t,f)\,\boldsymbol{y}(t,f)^H \tag{7}$$

$$\widehat{\Phi}_n(f) = \frac{1}{\sum_t \xi(t,f)}\sum_t \xi(t,f)\boldsymbol{y}(t,f)\boldsymbol{y}(t,f)^H \tag{8}$$

with $\eta(t,f)$ computed using Eq. (5) and $\xi(t,f)$ computed as

$$\xi(t,f) = \prod_{i=1}^{D}(1 - \widehat{M}_i(t,f)) \tag{9}$$

Essentially, Eq. (7) and (5) mean that only the T-F units dominated by speech are utilized to compute the speech covariance matrix, and the more speech-dominant a T-F unit is, the more weight is placed.

Next, following the free-field and plane-wave assumption [27], the unit-length steering vector for a potential location k is modeled as

$$\boldsymbol{c}(f,k) = \frac{1}{\sqrt{D}}\left[e^{-j2\pi\frac{f}{N}f_s\frac{d_{k1}}{c_s}}, \ldots, e^{-j2\pi\frac{f}{N}f_s\frac{d_{kD}}{c_s}}\right]^T, \tag{10}$$

where d_{ki} is the distance between sound source location k to microphone i and c_s is the speed of sound. Then, an MVDR beamformer is constructed:

$$\widehat{\boldsymbol{w}}(f,k) = \frac{\widehat{\Phi}_n(f)^{-1}\boldsymbol{c}(f,k)}{\boldsymbol{c}(f,k)^H\widehat{\Phi}_n(f)^{-1}\boldsymbol{c}(f,k)} \tag{11}$$

After that, the SNR of beamformed signal is computed from the energy of beamformed speech and beamformed noise.

$$\begin{aligned} &\mathrm{SNR}(k) \\ &= \sum_f \frac{\widehat{\boldsymbol{w}}(f,k)^{\boldsymbol{H}}\widehat{\Phi}_s(f)\widehat{\boldsymbol{w}}(f,k)}{\widehat{\boldsymbol{w}}(f,k)^{\boldsymbol{H}}\widehat{\Phi}_s(f)\widehat{\boldsymbol{w}}(f,k) + \widehat{\boldsymbol{w}}(f,k)^{\boldsymbol{H}}\widehat{\Phi}_n(f)\widehat{\boldsymbol{w}}(f,k)} \end{aligned} \tag{12}$$

Finally, the speaker location is predicted to be

$$\hat{k} = argmax_k\,\mathrm{SNR}(k) \tag{13}$$

Note that in Eq. (12), we constrain the SNR to be between zero and one. It is essentially similar to the PHAT weighting in the GCC-PHAT algorithm, where the GCC coefficient of each T-F unit is normalized to be between -1 and 1. In our experiments, we put more weights to higher-SNR frequencies:

$$\mathrm{SNR}(k) = \sum_f \frac{\gamma(f) * \hat{\boldsymbol{w}}(f,k)^{\boldsymbol{H}} \hat{\Phi}_s(f) \hat{\boldsymbol{w}}(f,k)}{\hat{\boldsymbol{w}}(f,k)^{\boldsymbol{H}} \hat{\Phi}_s(f) \hat{\boldsymbol{w}}(f,k) + \hat{\boldsymbol{w}}(f,k)^{\boldsymbol{H}} \hat{\Phi}_n(f) \hat{\boldsymbol{w}}(f,k)}, \quad (14)$$

where $\gamma(f)$ is defined as

$$\gamma(f) = \sum_t \eta(t,f) \quad (15)$$

Note that the sum of the combined speech mask within each frequency is used to indicate the importance of each frequency. In our experiments, much better results have been observed using Eq. (14) than using Eq. (12).

2.4. TDOA Estimation from Steering Vectors

Time-frequency masking based on deep learning has shown considerable potential for beamforming and robust ASR in the recent CHiME challenges [25], [24], [26], [28]. The major advance is to use the estimated masks from a powerful DNN to compute speech and noise statistics that are critical for accurate beamforming. Remarkable improvements have been observed on many robust ASR tasks [29]. In this context, one potential way to do robust TDOA estimation is to derive the time delay from the estimated steering vectors, as they contain sufficient information for robust TDOA estimation.

Following [24], [26], [30], the steering vector at each frequency is estimated using:

$$\hat{\boldsymbol{c}}(f) = \mathcal{P}\{\hat{\Phi}_s(f)\} = [\frac{1}{\sqrt{\hat{A}(f)^2+1}}, \frac{\hat{A}(f)}{\sqrt{\hat{A}(f)^2+1}} e^{j\hat{\theta}(f)}]^T, \quad (16)$$

where $\mathcal{P}\{\cdot\}$ extracts the principal eigenvector of the estimated speech covariance matrix computed in Eq. (7). Note that if $\hat{\Phi}_s(f)$ is well-estimated, it would be close to a rank-one matrix, and hence its principal eigenvector is a reasonable estimate of the steering vector [24], [27]. To derive the underlying time delay $\hat{\tau}$, we enumerate all the potential time delays and find the one that maximizes the following objective:

$$Sim(\tau) = \sum_f \gamma(f) cos\left(\hat{\theta}(f) - (-2\pi \frac{f}{N} f_s \tau)\right) \quad (17)$$

$$\hat{\tau} = argmax_\tau \, Sim(\tau) \quad (18)$$

The rationale is that the steering vector $\hat{\boldsymbol{c}}(f)$ is independently estimated at each frequency. Therefore, $\hat{\theta}(f)$ does not strictly follow the linear phase assumption. Our study enumerates all the potential time delays and finds a time delay τ with its phase delay $-2\pi \frac{f}{N} f_s \tau$ best matched with $\hat{\theta}(f)$ at every frequency as the final prediction. Following Eq. (14), we use $\gamma(f)$ in Eq. (15) to emphasize the frequencies with higher SNR. The cosine operation is naturally bounded, thus explicit normalization as in Eq. (3) and (12) is no longer necessary.

There are previous studies [31], [32], [33] trying to derive time delays from estimated steering vectors at each frequency or each T-F unit. They usually assume that there is no phase-wrapping, and divide the estimated phase delay by the angular frequency to get the time delay. This is however not realistic in practice, as there could be multiple time delays given exactly the same phase delay due to spatial aliasing and phase wrapping. Instead, our approach avoids this ambiguity by enumerating all the time delays and checking the similarity objective at each time delay. This strategy is more reasonable as a time delay is deterministically corresponding to a phase delay [34]. Another major difference between our approach and the previous studies is that we are using powerful DNNs for mask estimation, while previous studies are focused on using spatial clustering [31], [32] and empirical rules [33].

2.5. Mask Estimation

All the three proposed algorithms requires an accurate estimation of the speech mask $\hat{M}_i$. It is suggested in monaural speech separation [14], [13], [35] that DNNs are capable of accurately determining the speech dominance at each T-F unit. Among various types of neural networks, the bi-directional long short-term memory (BLSTM) [36] has been shown to produce consistently better separation results. In our study, a BLSTM is trained to estimate the ideal ratio mask. Depending on using the direct sound or the reverberant speech signal as the target, there are two ways to define the IRM:

$$\mathrm{IRM}_i^{reverb}(t,f) = \frac{|c_i(f)s(t,f) + h_i(t,f)|^2}{|c_i(f)s(t,f) + h_i(t,f)|^2 + |n_i(t,f)|^2} \quad (19)$$

$$\mathrm{IRM}_i^{direct}(t,f) = \frac{|c_i(f)s(t,f)|^2}{|c_i(f)s(t,f)|^2 + |h_i(t,f) + n_i(t,f)|^2} \quad (20)$$

where i indexes the microphone channel.

3. Experimental Setup

The proposed algorithms are evaluated using a binaural setup and a two-microphone setup for robust TDOA estimation in highly reverberant environments with strong diffuse babble noise. Fig. 1 depicts the experimental setup.

In the binaural setup, the simulated binaural room impulse responses[1] (BRIR) with T60 ranging from 0.0s to 1.0s in steps of 0.1s generated using the CATT software is used for training. The simulated room size is fixed at 6x4x3m. The BRIRs are measured with the array placed around the center of the room and at a height of 2m, and the source located at one of the 37 directions (from -90° to 90° in steps of 5°) at the same height as the array and at a distance of 1.5m to the array center. The real BRIRs[2] captured using a HATS dummy head in four real rooms with different sizes and T60s are used for testing. The dummy head is placed at a height of 2.8m, and the source to array distance is 1.5m. The real BRIRs are also measured using the same 37 directions. We put 37 different interference speakers at each of the 37 directions and the target speaker at one of the directions. The 720 IEEE female utterances are utilized as the target speech 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 concatenate the utterances of each of the 630 speakers in the TIMIT dataset and randomly pick 37 speech segments from 37 randomly-chosen speakers to put at each of the 37 directions. For each speaker in the babble noise, we use the first half of the concatenated utterance to generate the training and validation noise and the second half to generate the testing noise. There are 10,000, 800 and 3,000 binaural mixtures in the training, validation and testing set.

In the two-microphone setup, the RIR generator[3] based on the image method is employed to generate the RIRs. For the training and validation data, we put one interference speaker at each of the 36 directions spanning 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 ranging from $-90°$ to $90°$ in steps of $5°$, and the target speaker at one of the 37 directions. This way, the testing RIRs are unseen during training. The distance between each speaker and the array center is 1m. The room size is fixed at 8x8x3m, and the two microphones are placed around the

[1] Available at http://iosr.uk/software/index.php#CATT_RIRs.
[2] Available at https://github.com/IoSR-Surrey/RealRoomBRIRs.
[3] Available at https://github.com/ehabets/RIR-Generator.

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

IRM Type	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
(19)	Mask-Weighted GCC-PHAT	78.5	95.5	97.6	94.0	90.3	84.7	81.3	68.6	62.3	57.9	53.3
	Using IRM	98.0	100.0	100.0	100.0	100.0	99.7	98.6	97.7	96.1	93.0	94.6
	Mask-Weighted Steered-Response SNR	86.7	97.4	95.1	94.4	91.3	90.1	89.3	82.3	79.7	75.6	72.1
	Using IRM	99.6	100.0	100.0	99.7	100.0	99.7	100.0	99.7	99.4	98.0	99.3
	TDOA estimation from Steering Vectors	86.4	96.2	96.5	96.3	93.4	89.1	86.2	82.0	79.7	75.9	68.1
	Using IRM to get oracle $\hat{\boldsymbol{c}}(f)$	98.5	100.0	100.0	100.0	100.0	100.0	99.3	97.7	96.8	95.3	96.4
(20)	Mask-Weighted GCC-PHAT	88.2	97.1	96.5	95.7	94.1	91.5	89.3	83.4	80.6	79.6	73.9
	Using IRM	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.7	100.0
	Mask-Weighted Steered-Response SNR	90.5	98.1	95.5	95.0	94.1	95.6	93.8	87.5	85.2	81.6	78.6
	Using IRM	99.9	100.0	100.0	99.7	100.0	100.0	100.0	99.7	100.0	100.0	100.0
	TDOA Estimation from Steering Vectors	**91.0**	96.8	95.5	97.0	95.8	93.5	91.7	87.8	86.8	84.3	80.8
	Using IRM to get oracle $\hat{\boldsymbol{c}}(f)$	99.8	100.0	100.0	100.0	100.0	100.0	100.0	99.7	100.0	98.7	100.0

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 IEEE and TIMIT utterances are used to generate the same number of training, validation and testing utterances in the same way as in the binaural setup.

The average duration of the mixture is 2.4s. The input SNR computed from reverberant speech and reverberant noise for both dataset is -6dB. Note that if we consider the direct sound signal as target speech and the rest as noise, the SNR would be even lower. We train our BLSTM using all the single-channel signals (10,000*2 in total) in the training data. In the microphone array setup, log power spectrogram is used as the input feature, while in the binaural setup, interaural level differences (ILD) are also used. Sentence-level mean normalization is performed on the input features before global mean-variance normalization. The BLSTM contains two hidden layers each with 384 units in each direction. Sigmoidal units are used in the output layer. The Adam algorithm is used to minimize the mean squared error for mask estimation. The window size is 32ms and the hop size is 8ms. The sampling rate is 16 kHz. 512-point FFT is performed to extract 257-dimensional log spectrogram features at each frame.

We measure the performance in terms of gross accuracy, which considers a prediction is correct if the prediction is within 5^{o} (inclusive) from the true target direction.

Fig. 1. Illustration of (a) binaural, and (b) two-microphone setup.

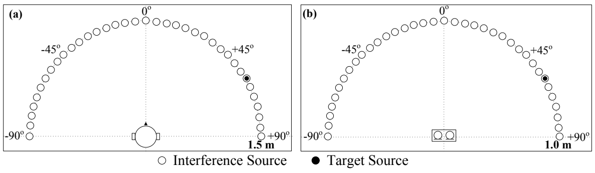

Table 2. Comparison of TDOA estimation performance (% Gross Accuracy) of different approaches in binaural setup.

IRM Type	Approaches	AVG	Room - T60(s)/DRR(dB)			
			A	B	C	D
			0.32/6.1	0.47/5.3	0.68/8.8	0.89/6.1
-	GCC-PHAT	29.4	27.3	30.7	35.9	23.8
(19)	Mask-Weighted GCC-PHAT	91.3	94.1	87.8	91.4	92.0
	Using IRM	99.1	98.8	99.8	98.9	98.8
	Mask-Weighted Steered-Response SNR	86.4	89.5	86.4	81.2	88.5
	Using IRM	99.4	99.0	100.0	99.2	99.5
	TDOA estimation from Steering Vectors	**92.0**	93.7	90.4	90.5	93.4
	Using IRM to get oracle $\hat{\boldsymbol{c}}(f)$	98.9	98.8	99.5	98.9	98.3
(20)	Mask-Weighted GCC-PHAT	90.8	94.2	87.7	89.8	91.5
	Using IRM	99.4	99.2	99.8	98.9	99.7
	Mask-Weighted Steered-Response SNR	70.0	79.2	65.5	70.3	64.9
	Using IRM	99.4	98.8	100	98.9	99.8
	TDOA Estimation from Steering Vectors	91.1	93.2	89.9	89.4	91.8
	Using IRM to get oracle $\hat{\boldsymbol{c}}(f)$	99.4	99.2	99.7	99.2	99.5

4. Evaluation Results

The gross accuracy results are reported in Table 1 and 2, together with the oracle performance marked in grey and the direction-to-reverberant energy ratio (DRR) at each T60 level. Using the estimated masks from the BLSTM, the proposed mask-weighted GCC-PHAT algorithm substantially improves the conventional GCC-PHAT algorithm (from 25.8% to 78.5% and 88.2% in Table 1, and from 29.4% to 91.3% and 90.8% in Table 2). The TDOA estimation algorithm based on steering vectors exhibits the strongest robustness among all the competing algorithms, especially when the T60 is high. Interestingly, using the IRM defined using the direct sound as target leads to almost 100% gross accuracy for all the proposed algorithms (100.0%, 99.9% and 99.8% in Table 1, and 99.4%, 99.4% and 99.4% in Table 2), indicating the strong potential of the T-F masking based approach for robust TDOA estimation. Using the reverberant speech as target is only slightly worse in the oracle case (98.0%, 99.6% and 98.5% in Table 1, and 99.1%, 99.4% and 98.9% in Table 2). In the two-microphone setup, estimating the IRM defined using the direct sound as target leads to consistently better performance over the alternative definition (88.2% v.s. 78.5%, 90.5% v.s. 87.7%, and 91.0% v.s. 86.4%). This makes sense as the time delay information is mostly contained in the direct sound. However, estimating the IRM defined using the alternative definition gets slightly better performance in the binaural setup (91.3% v.s. 90.8%, 86.4% v.s. 70.0%, and 92.0% v.s. 91.1%). This is possibly due to the head shadowing effects and the mismatch between training and testing BRIRs in the binaural setup. The mask-weighted steered-response SNR algorithm performs less impressive in the binaural setup than in the two-microphone setup, likely because the gains at different channels cannot be simply treated as equal in the binaural case considering the head shadowing effects.

5. Concluding Remarks

We have proposed three novel algorithms based on T-F masking and deep learning for robust TDOA estimation. All of them show strong robustness in environments with low SNR and high reverberation. Future research would extend the algorithms to multi-channel cases, modify it to deal with moving sound sources, and evaluate its potential for beamforming and robust ASR. Before closing, we emphasize again that the proposed algorithms formulate robust speaker localization as a T-F mask estimation problem. It has much more potential over conventional algorithms, as now robust speaker localization can be approached and benefited from deep learning, a rapidly advancing field.

6. References

[1] 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.

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

[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] J.-M. Valin, F. Michaud, J. Rouat, and D. Letourneau, "Robust Sound Source Localization using A Microphone Array on A Mobile Robot," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003, vol. 2, pp. 1228–1233.

[5] Y. Rui and D. Florencio, "Time Delay Estimation in the Presence of Correlated Noise and Reverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2004, pp. 133–136.

[6] H.-G. Kang, M. Graczyk, and J. Skoglund, "On Pre-Filtering Strategies for the GCC-PHAT Algorithm," in *International Workshop on Acoustic Signal Enhancement*, 2016, pp. 1–5.

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

[8] J. Woodruff and D. L. Wang, "Binaural Localization of Multiple Sources in Reverberant and Noisy Environments," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 5, pp. 1503–1512, 2012.

[9] J. Woodruff and D. L. Wang, "Binaural Detection, Localization, and Segregation in Reverberant Environments based on Joint Pitch and Azimuth Cues," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 4, pp. 806–815, 2013.

[10] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An Algorithm that Improves Speech Intelligibility in Noise for Normal-Hearing Listeners," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.

[11] X. Xiao, S. Zhao, T. N. T. Nguyen, D. L. Jones, E. S. Chng, and H. Li, "An Expectation-Maximization Eigenvector Clustering Approach to Direction of Arrival Estimation of Multiple Speech Sources," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 6330–6334.

[12] D. L. Wang, "Deep Learning Reinvents the Hearing Aid," *IEEE Spectrum*, vol. 54, no. 3, pp. 32–37, 2017.

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

[14] Y. Wang and D.L. Wang, "Towards Scaling Up Classification-Based Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[15] 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.

[16] P. Pertila and E. Cakir, "Robust Direction Estimation with Convolutional Neural Networks based Steered Response Power," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 6125–6129.

[17] C. Xu, X. Xiao, S. Sun, W. Rao, E. S. Chng, and H. Li, "Weighted Spatial Covariance Matrix Estimation for MUSIC based TDOA Estimation of Speech Source," in *Proceedings of Interspeech*, 2017, pp. 1894–1898.

[18] N. Ma, G. Brown, and T. May, "Exploiting Deep Neural Networks and Head Movements for Binaural Localisation of Multiple Speakers in Reverberant Conditions," in *Proceedings of Interspeech*, 2015, pp. 160–164.

[19] S. Chakrabarty and E. A. P. Habets, "Broadband DOA Estimation using Convolutional Neural Networks Trained with Noise Signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2017, pp. 136–140.

[20] C. Zhang, D. Florêncio, and Z. Zhang, "Why Does PHAT Work Well in Lownoise, Reverberative Environments?," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 2565–2568.

[21] J. M. Valin, F. Michaud, and J. Rouat, "Robust Localization and Tracking of Simultaneous Moving Sound Sources using Beamforming and Particle Filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216–228, 2007.

[22] F. Grondin and F. Michaud, "Time Difference of Arrival Estimation based on Binary Frequency Mask for Sound Source Localization on Mobile Robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015, pp. 6149–6154.

[23] C. Blandin, A. Ozerov, and E. Vincent, "Multi-Source TDOA Estimation in Reverberant Audio using Angular Spectra and Clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-Based Expectation-Maximization Source Separation and Localization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[35] 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.

[36] 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.