Robust Speaker Localization Guided by Deep Learning-Based Time-Frequency Masking

Zhong-Qiu Wang, Xueliang Zhang, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Deep learning-based time-frequency (T-F) masking has dramatically advanced monaural (single-channel) speech separation and enhancement. This study investigates its potential for direction of arrival (DOA) estimation in noisy and reverberant environments. We explore ways of combining T-F masking and conventional localization algorithms, such as generalized cross correlation with phase transform, as well as newly proposed algorithms based on steered-response SNR and steering vectors. The key idea is to utilize deep neural networks (DNNs) to identify speech dominant T-F units containing relatively clean phase for DOA estimation. Our DNN is trained using only monaural spectral information, and this makes the trained model directly applicable to arrays with various numbers of microphones arranged in diverse geometries. Although only monaural information is used for training, experimental results show strong robustness of the proposed approach in new environments with intense noise and room reverberation, outperforming traditional DOA estimation methods by large margins. Our study also suggests that the ideal ratio mask and its variants remain effective training targets for robust speaker localization.**

***Index Terms*—GCC-PHAT, steered-response power, time-frequency masking, robust speaker localization, deep neural networks.**

I. Introduction

ROBUST speaker localization has many applications in real-world tasks, such as teleconferencing, robotics and voice-activated human-computer interaction. For example, the ability to localize a speaker in daily environments is important for a voice-based interface such as Amazon Echo. Localization is also widely used in beamforming for speech separation or enhancement [1]. Conventionally, generalized cross correlation with phase transform (GCC-PHAT) [2] (or steered-response power with phase transform (SRP-PHAT) [3]) and multiple signal classification (MUSIC) [4] are the two most popular algorithms for sound source localization, both originating from narrowband antenna signal processing. However, their speaker localization performance is unsatisfactory in noisy and reverberant environments; in such environments, the summation of GCC coefficients exhibits spurious peaks and the noise subspace constructed in the MUSIC algorithm does not correspond to the true noise subspace.

To improve the robustness to noise and reverberation, frequency-dependent SNR (signal-to-noise ratio) weighting is designed to emphasize frequencies with higher SNR for the GCC-PHAT algorithm. SNR can be computed in various ways, such as rule-based methods [5], voice activity detection based algorithms [6], or minimum mean square error based approaches [7]. T-F unit level SNR based on minima controlled recursive averaging or inter-channel coherence has also been applied to emphasize T-F units with higher SNR or coherence [8]–[10]. However, these algorithms typically assume stationary noise, which is an unrealistic assumption in real-world acoustic environments.

While it is difficult to perform multi-channel localization in noisy and reverberant environments, with two ears the human auditory system shows a remarkable capacity at localizing sound sources. Psychoacoustic evidence suggests that sound localization largely depends on sound separation [11]–[13], which operates according to auditory scene analysis principles [11]. Motivated by perceptual organization, we approach robust speaker localization from the angle of monaural speech separation.

It is well-known that, even for a severely corrupted utterance, there are still many T-F units dominated by target speech [13]. As analyzed in [9], [14]–[18], these T-F units carry relatively clean phase and may be sufficient for speaker localization. Motivated by this observation, our approach aims at identifying speech dominant T-F units at each microphone channel and only using such T-F units for multi-channel localization. A profound consequence of this new approach is that deep learning can be brought to bear on T-F unit level classification or regression for robust localization. Recently, deep learning based time-frequency masking has dramatically elevated monaural speech separation and enhancement performance (see [19] for an overview). Thanks to the strong learning capacity of deep neural networks, they can accurately determine the speech or noise dominance at each T-F unit [20].

In this context, we perform robust DOA estimation by utilizing deep learning based T-F masking. This study makes three contributions. First, DNN estimated masks

Manuscript received February 9, 2018; revised July 27, 2018 and September 19, 2018; accepted October 11, 2018. Date of publication October 15, 2018; date of current version October 29, 2018. This work was supported in part by the Air Force Research Laboratory contract (FA8750-15-1-0279), in part by the National Science Foundation grant (IIS-1409431), and in part by the Ohio Supercomputer Center. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Huseyin Hacihabiboglu. *(Corresponding author: Zhong-Qiu Wang.)*

Z.-Q. Wang is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: wangzhon@cse.ohio-state.edu).

X. Zhang is with the Department of Computer Science, Inner Mongolia University, Hohhot 010021, China (e-mail: cszxl@imu.edu.cn).

D. Wang is with the Department of Computer Science and Engineering, and the Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210 USA (e-mail: dwang@cse.ohio-state.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2018.2876169

are utilized to improve the robustness of conventional cross-correlation-based, beamforming-based and subspace-based algorithms [3] for DOA estimation in environments with strong noise and reverberation, following previous research along similar directions [21], [22]. A key ingredient, we believe, is balancing the contributions of individual frequency bands for the DOA estimation of broadband speech signals. Second, we find that using the IRM and its variants, which consider direct sound as the target signal, leads to high localization accuracy, suggesting that such training targets are very effective for robust speaker localization (see also [21]). Third, we show that the trained model is versatile in application to sensor arrays with diverse geometries and with various numbers of microphones.

The rest of this paper is organized as follows. The proposed algorithms are presented in Section II. Experimental setup and evaluation results are reported in Sections III and IV. Section V concludes this paper. Note that a preliminary version of this work has been recently accepted by Interspeech [23]. The present study extends the preliminary work on time delay estimation in the two-sensor case to multi-sensor arrays with arbitrary array geometry.

II. System Description

We start with a review of the classic GCC-PHAT algorithm, which motivated the design of our algorithms. The three subsequent sections present three proposed localization algorithms based on mask-weighted GCC-PHAT, mask-weighted steered-response SNR, and steering vectors. These proposed algorithms respectively represent the cross-correlation-, beamforming- and subspace-based approaches for localization. Deep learning based time-frequency masking is described in the last section.

Suppose that there is only one target speaker, the physical model for a pair of signals in noisy and reverberant environments under the narrowband approximation assumption can be formulated as

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

where $s(t,f)$ is the STFT (short-time Fourier transform) value of the direct-path signal of the target speaker captured by a reference microphone at time t and frequency f, and $\boldsymbol{c}(f)$ is the relative transfer function. $\boldsymbol{c}(f)s(t,f)$, $\boldsymbol{h}(t,f)$, $\boldsymbol{n}(t,f)$, and $\boldsymbol{y}(t,f)$ represent the STFT vectors of the direct signal, its reverberation, reverberated noise, and received mixture, respectively. By designating the first microphone as the reference, the relative transfer function, $\boldsymbol{c}(f)$, can be described as

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

where τ^* denotes the time difference of arrival (TDOA) between the two signals in seconds, $A(f)$ is a real-valued relative gain, j is the imaginary unit, f_s is the sampling rate in Hz, N is the number of discrete Fourier transform (DFT) frequencies, and $[\cdot]^T$ stands for transpose. Note that the range of f is from 0 to $N/2$.

The classical GCC-PHAT algorithm [2], [3] estimates the time delay of a pair of microphones p and q by computing their generalized cross-correlation coefficients with a weighting mechanism based on phase transform

$$GCC_{p,q}\,(t,f,k) = \mathcal{R}e\left\{\frac{y_p(t,f)y_q(t,f)^H}{|y_p(t,f)||y_q(t,f)^H|}e^{-j2\pi\frac{f}{N}f_s\tau_{p,q}(k)}\right\}$$

$$= \cos\left(\angle y_p\,(t,f) - \angle y_q\,(t,f) - 2\pi\frac{f}{N}f_s\tau_{p,q}\,(k)\right) \qquad (3)$$

where $(\cdot)^H$ represents conjugate transpose, $\mathcal{R}e\{\cdot\}$ extracts the real part, $|\cdot|$ computes the magnitude, and $\angle(\cdot)$ extracts the phase. $\tau_{p,q}\ (k) = (d_{kq} - d_{kp})/c_s$ denotes the time delay of a candidate direction or location k, where c_s is the speed of sound in the air, and d_{kq} and d_{kp} represent the distance between the hypothesized sound source to microphone p and q, respectively. Assuming that the target speaker is fixed within a single utterance, the GCC coefficients are then summated and the time delay producing the largest summation represents the delay estimate.

Intuitively, this algorithm first aligns two microphone signals using a candidate time delay τ and then computes their cosine distance at each T-F unit pair. If the cosine distance is close to one, it means that the candidate time delay is close to the true time delay at that T-F unit. The summation functions as a voting mechanism to combine the observations at all the unit pairs. Since each GCC coefficient is naturally bounded between -1 and 1, each T-F unit pair has an equal contribution to the summation. Note that PHAT weighting [24], [25], i.e., the magnitude normalization term in Eq. (3), is critical here, as the energy of human speech is mostly concentrated in lower frequency bands. If the magnitude normalization is not performed, lower frequency components would have much larger GCC coefficients and dominate the summation, making it less sharp. In addition, the scales of the two signals are usually different in near-field or binaural cases. It is hence beneficial to remove the influence of different energy levels.

We emphasize that summation over frequencies in the GCC-PHAT algorithm is very important for broadband speech signals. Because of spatial aliasing [1], the cross-correlation function at high frequencies is typically periodic, containing multiple peaks. It is hence important to summate over all the frequencies to sharpen the peak corresponding to the true timed delay [13].

Although GCC-PHAT performs well in environments with low to moderate reverberation, it is susceptible to strong reverberation and noise. To see this, suppose that there is a strong directional noise source. There would be many T-F units dominated by the noise source. In this case, the noise source would exhibit the highest peak in the summated GCC coefficients. Similarly, diffuse noise and reverberation would broaden GCC peaks and corrupt TDOA estimation.

A. Mask-Weighted GCC-PHAT

The time delay information is contained in the direct sound signal, $\boldsymbol{c}(f)s(t,f)$. Including the GCC coefficients of any T-F unit pairs dominated by noise or reverberation in the summation would weaken localization performance. To improve robustness, we multiply the GCC coefficients for a pair of microphones and

a masking-based weighting term following [9], [17]:

$$MGCC_{p,q}\ (t,f,k) = M_{p,q}^{(s)}\ (t,f)\, GCC_{p,q}\left(t,f,\tau_{p,q}\left(k\right)\right) \tag{4}$$

where $M_{p,q}^{(s)}(t,f)$ represents the importance of the T-F unit pair for TDOA estimation (superscript (s) indicates target signal – see Eq. (1)). It is computed using

$$M_{p,q}^{(s)}\left(t,f\right) = M_p\left(t,f\right) M_q\left(t,f\right) \tag{5}$$

where M_p and M_q denote the T-F masks representing the estimated speech portion at each T-F unit of microphone p and q, respectively. The estimated masks should be close to one for T-F units dominated by direct sound signals and zero for T-F units dominated by noise or reverberation. Mask estimation based on deep learning will be discussed later in Section II-E. The time delay or direction is then computed as

$$\hat{k} = \arg\max_{k} \sum_{(p,q)\epsilon\Omega} \sum_{t} \sum_{f=1}^{N/2} MGCC_{p,q}\left(t,f,k\right) \tag{6}$$

where Ω represents the set of microphones pairs in an array used for the summation. Note that the above delay estimation is formulated for a general array with at least two sensors.

Through the product of the masks of individual microphone channels, the weighting mechanism in Eq. (5) places more weights on the T-F units dominated by target speech across all the microphone channels. This makes sense as target-dominant T-F units carry cleaner phase information for localization than other ones. Therefore, adding this weighting term should sharpen the peak corresponding to the target source in the summation and suppress the peaks corresponding to noise sources and reverberation.

At a conceptual level, T-F masking guides localization in the following sense. First, T-F masking serves to specify what the target source is through supervised training. Although we are interested in speaker localization in this study, the framework does not change if one is interested in localizing, for example, musical instruments instead. Second, masking suppresses the impact of interfering sounds and reverberation in localization. Without masking's guidance, traditional DOA estimation could be considered "blind" as it is indiscriminately based on sound energy in one form or another.

One property of the proposed algorithm is that, for relatively clean utterances, estimated mask values would all be close to one. In such a case, the proposed algorithm simply reduces to the classic GCC-PHAT algorithm, which is known to perform very well in clean environments [3].

We point out that our approach is different from applying the GCC-PHAT algorithm to enhanced speech signals obtained via T-F masking. To explain this, let us substitute $M_p(t,f)y_p(t,f)$ and $M_q(t,f)y_q(t,f)$ for $y_p(t,f)$ and $y_q(t,f)$ in Eq. (3). Doing it this way produces the same GCC coefficients as using the unprocessed $y_p(t,f)$ and $y_q(t,f)$, because the real-valued masks are cancelled out due to the PHAT weighting (unless time-domain re-synthesis is performed). The proposed algorithm utilizes estimated masks as a weighting mechanism to identify for localization speech dominant T-F units where the phase information is less contaminated. Note that localization cues are mostly contained in inter-channel phase differences.

Our study first estimates a T-F mask for each single-channel signal and then combines the estimated masks using their product. In this way, the resulting DNN for mask estimation can be readily applied to microphone arrays with various numbers of microphones arranged in arbitrary geometry, although geometrical information is still necessary for DOA estimation. This flexibility distinguishes our algorithms from classification based approaches [26]–[30] for DOA estimation, which typically require fixed microphone geometry, fixed number of microphones and fixed spatial resolution for DNN training and testing. In addition, the trained neural network for mask estimation can be directly employed for related tasks such as voice activity detection, spatial covariance matrix estimation, beamforming, and single-channel post-filtering [31], [32]. Also, classification approaches usually rely heavily on spatial information for DNN training and therefore may not work well when strong directional noise sources are present. In contrast, our approach suppresses the peaks corresponding to interfering sources via T-F masking, which is trained to separate a target source.

Following [9], [17], a recent study [21] proposed to use DNN based T-F masking to improve the SRP-PHAT algorithm. This method first averages the log-magnitudes from all the channels and then uses a convolutional neural network to estimate an average mask from the averaged magnitudes. The estimated average mask is then used as weights for the SRP-PHAT algorithm. Averaging log-magnitudes would not be a good idea when the signals at different channels vary significantly, for example in the binaural case where interaural level differences can be large. In addition, averaging would incorporate contaminated T-F units for DOA estimation. In contrast, our approach estimates a mask from each microphone signal separately, using features extracted from that microphone. We then combine estimated masks using the product rule in Eq. (5). As a result, our approach places more weights on the T-F units dominated by target speech in all the microphone channels. It should, however, be noted that performing channel-wise mask estimation comes at the cost of increased computation compared to estimating an average mask. Furthermore, as described in Section II-E, our study uses powerful recurrent neural networks (RNNs) to estimate the IRM [33] and phase-sensitive mask [34], [35], yielding better mask estimation for localization.

B. Mask-Weighted Steered-Response SNR

The GCC-PHAT, SRP-PHAT or BeamScan [36], [37] algorithms steer a beam towards a hypothesized direction and compute the steered-response power of noisy speech to determine whether the hypothesized direction is the target direction, i.e., with the strongest response. The proposed mask-weighted GCC-PHAT algorithm utilizes a time-frequency mask to emphasize speech dominant T-F units so that the steered-response power of estimated target speech, rather than noisy speech, is used as the location indicator. This section uses steered-response SNR as the indicator, as the SNR considers both speech power and noise power, and more importantly, the SNR at each frequency can be bounded between zero and one so that DOA estimation

would not be biased towards high-energy lower frequency components. Specifically, for each direction of interest, we design a beamformer to point towards that direction, and the direction producing the highest SNR is considered as the predicted target direction [10]. Speech and noise covariance matrices for beamforming and SNR computation can be robustly estimated with the guidance of T-F masking.

Let $\boldsymbol{y}_{p,q}\ (t,f) = [y_p(t,f), y_q(t,f)]^T$. The speech and noise covariance matrices between microphone p and q at each frequency are computed in the following way,

$$\hat{\Phi}^{(s)}_{p,q}\ (f) = \frac{\sum_t M^{(s)}_{p,q}\ (t,f)\, \boldsymbol{y}_{p,q}\,(t,f)\ \boldsymbol{y}_{p,q}(t,f)^H}{\sum_t M^{(s)}_{p,q}\ (t,f)} \tag{7}$$

$$\hat{\Phi}^{(n)}_{p,q}\ (f) = \frac{\sum_t M^{(n)}_{p,q}\ (t,f)\, \boldsymbol{y}_{p,q}\,(t,f)\, \boldsymbol{y}_{p,q}(t,f)^H}{\sum_t M^{(n)}_{p,q}\ (t,f)} \tag{8}$$

where $M^{(s)}_{p,q}(t,f)$ is given in Eq. (5) and $M^{(n)}_{p,q}(t,f)$ is computed as (superscript (n) indicates noise or interference)

$$M^{(n)}_{p,q}\ (t,f)\ = (1 - M_p\ (t,f))\ (1 - M_q\ (t,f)) \tag{9}$$

Motivated by the work in masking-based beamforming for automatic speech recognition (ASR) [38], [31] (see also [32]), the weights in Eq. (7) are empirically designed so that only the T-F units dominated by speech in both microphone channels are utilized to compute the speech covariance matrix, and the more speech-dominant a T-F unit is, the more weight is placed on it. The noise covariance matrix is computed in a similar fashion, where the noise mask is simply obtained in (9) as the complement of the speech mask.

Next, under the plane-wave and far-field assumption [1], the steering vector for a candidate direction k is modeled as

$$\boldsymbol{c}_{p,q}\ (f,k) = \left[e^{-j2\pi\frac{f}{N}f_s\frac{d_{kp}}{cs}}, e^{-j2\pi\frac{f}{N}f_s\frac{d_{kq}}{cs}}\right]^T \tag{10}$$

Then, $\boldsymbol{c}_{p,q}(f,k)$ is normalized to unit length,

$$\bar{\boldsymbol{c}}_{p,q}\ (f,k) = \frac{\boldsymbol{c}_{p,q}\,(f,k)}{\|\boldsymbol{c}_{p,q}\,(f,k)\|} \tag{11}$$

and a minimum variance distortion-less response (MVDR) beamformer is constructed:

$$\boldsymbol{w}_{p,q}\ (f,k) = \frac{\hat{\Phi}^{(n)}_{p,q}(f)^{-1}\bar{\boldsymbol{c}}_{p,q}}{\bar{\boldsymbol{c}}_{p,q}{}^H\hat{\Phi}^{(n)}_{p,q}(f)^{-1}\bar{\boldsymbol{c}}_{p,q}} \tag{12}$$

Afterwards, the SNR of the beamformed signal is estimated as the ratio between the beamformed speech energy and beamformed noise energy.

$$\mathrm{SNR}_{p,q}\ (f,k) = \frac{\boldsymbol{w}_{p,q}(f,k)^H\hat{\Phi}^{(s)}_{p,q}\,(f)\,\boldsymbol{w}_{p,q}\,(f,k)}{\boldsymbol{w}_{p,q}(f,k)^H\hat{\Phi}^{(n)}_{p,q}\,(f)\,\boldsymbol{w}_{p,q}\,(f,k)} \tag{13}$$

Finally, the speaker location is estimated as

$$\hat{k} = \arg\max_k \sum_{(p,q)\epsilon\Omega}\sum_{f=1}^{N/2}\mathrm{SNR}_{p,q}\,(f,k) \tag{14}$$

One issue with Eq. (13) is that the computed energy and SNR are unbounded at each frequency band. In such cases, several frequency bands may dominate the SNR calculation. To avoid this problem, we restrict it to between zero and one in the following way.

$$\mathrm{SNR}_{p,q}\ (f,k) = \frac{\boldsymbol{w}_{p,q}(f,k)^H\hat{\Phi}^{(s)}_{p,q}(f)\boldsymbol{w}_{p,q}(f,k)}{\boldsymbol{w}_{p,q}(f,k)^H\hat{\Phi}^{(s)}_{p,q}(f)\boldsymbol{w}_{p,q}(f,k)+\boldsymbol{w}_{p,q}(f,k)^H\hat{\Phi}^{(n)}_{p,q}(f)\boldsymbol{w}_{p,q}(f,k)} \tag{15}$$

Eq. (15) shares the same spirit as PHAT weighting, where the GCC coefficient at each unit pair is bounded between -1 and 1, making each frequency contribute equally to the summation.

One can also explore alternative ways of weighting different frequency bands. One of them is to place more weights on higher-SNR frequency bands, i.e.,

$$\mathrm{SNR}_{p,q}\ (f,k) = \frac{\bar{M}_{p,q}(f)\ \boldsymbol{w}_{p,q}(f,k)^H\hat{\Phi}^{(s)}_{p,q}(f)\boldsymbol{w}_{p,q}(f,k)}{\boldsymbol{w}_{p,q}(f,k)^H\hat{\Phi}^{(s)}_{p,q}(f)\boldsymbol{w}_{p,q}(f,k)\ +\ \boldsymbol{w}_{p,q}(f,k)^H\hat{\Phi}^{(n)}_{p,q}(f)\boldsymbol{w}_{p,q}(f,k)} \tag{16}$$

$$\bar{M}_{p,q}\ (f) = \sum_t M^{(s)}_{p,q}\,(t,f)\,/\sum_{t,f} M^{(s)}_{p,q}\,(t,f) \tag{17}$$

where the sum of the speech mask $M^{(s)}_{p,q}(t,f)$ within each frequency band is used to indicate the importance of that band for localization. This frequency weighting, which counters the energy normalization, is motivated by the mask-weighted GCC-PHAT algorithm, which implicitly places more weights on frequencies with larger $\bar{M}_{p,q}(f)$. In our experiments, consistently better performance is observed using Eq. (16) than using Eq. (13) and (15) (see Section IV).

C. DOA Estimation Based on Steering Vectors

In the recent CHiME-3 and 4 challenges [39], [40], deep learning based time-frequency masking has been prominently employed for acoustic beamforming and robust ASR [31], [38], [32]. The main idea is to utilize estimated masks to compute the spatial covariance matrices and steering vectors that are critical for accurate beamforming. Remarkable improvements in terms of ASR performance have been reported over conventional beamforming techniques that employ traditional DOA estimation algorithms such as GCC-PHAT [41] and SRP-PHAT [39] for steering vector computation. This success is largely attributed to the power of deep learning based mask estimation [19]. In this context, we propose to perform DOA estimation from estimated steering vectors, as they contain sufficient information about the underlying target direction.

Following [38], [32], the steering vector for microphone p and q, $\hat{\boldsymbol{c}}_{p,q}(f)$, is estimated as the principal eigenvector of the estimated speech covariance matrix computed using Eq. (7). If $\hat{\Phi}^{(s)}_{p,q}(f)$ is accurately estimated, it would be close to a rank-one matrix, as the target speaker is a directional source and its principal eigenvector is a reasonable estimate of the steering vector [1].

To derive the underlying time delay or direction, we enumerate all the candidate directions and find the direction that maximizes the following similarity:

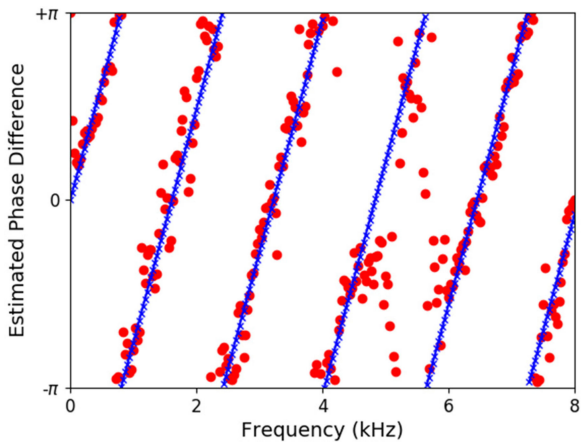

Fig. 1. Illustration of DOA estimation based on estimated steering vectors for a 2.4 s two-microphone (spacing: 24 cm) signal with babble noise. The SNR level is −6 dB and reverberation time is 0.16 s. Dots indicate the estimated phase differences $\angle(\hat{c}_{p,q}(f))_1 - \angle(\hat{c}_{p,q}(f))_2$ obtained using the IRM, and crosses the fitted phase differences $2\pi \frac{f}{N} f_s \tau_{p,q}(k)$ for a candidate direction k at each frequency.

$$S_{p,q}\ (f,k) = cos\left(\angle(\hat{c}_{p,q}\,(f))_1 - \angle(\hat{c}_{p,q}\,(f))_2 - 2\pi\frac{f}{N} f_s \tau_{p,q}\,(k)\right) \quad (18)$$

$$\hat{k} = \arg\max_k \sum_{(p,q)\epsilon\Omega} \sum_{f=1}^{N/2} S_{p,q}\,(f,k) \quad (19)$$

The rationale is that $\hat{\boldsymbol{c}}_{p,q}(f)$ is independently estimated at each frequency, and therefore the estimated phase difference, $\angle(\hat{c}_{p,q}(f))_1 - \angle(\hat{c}_{p,q}(f))_2$, between the two complex values in $\hat{\boldsymbol{c}}_{p,q}(f)$ does not strictly follow the linear phase assumption. We enumerate all the candidate directions and find as the final estimate a direction k with its hypothesized phase delay $2\pi \frac{f}{N} f_s \tau_{p,q}(k)$ that best matches the estimated phase difference at every frequency band. As illustrated in Fig. 1, this approach can be understood as performing circular linear regression between the estimated phase difference and frequency index f, where the slope is determined by $\tau_{p,q}(k)$ and the periodic cosine operation is employed to deal with phase wrapping. The cosine operation is naturally bounded between -1 and 1, thus explicit energy normalization as in Eq. (3) and (15) is not necessary. When there are more than two microphones, we simply combine all the microphone pairs by the summation. We optimize the similarity function through explicit enumeration. Eq. (18) in form is similar to Eq. (3). The key difference is that the phase difference per frequency is obtained from robustly estimated steering vectors rather than from the observed phase difference at each unit pair.

Similar to Eq. (16), we emphasize the frequency bands with higher SNR using $\bar{M}_{p,q}(f)$ given in Eq. (17).

$$S_{p,q}\,(f,k) = \bar{M}_{p,q}\,(f)\, cos\Big(\angle(\hat{c}_{p,q}\,(f))_1 - \angle(\hat{c}_{p,q}\,(f))_2 - 2\pi\frac{f}{N} f_s \tau_{p,q}\,(k)\Big) \quad (20)$$

Note that this algorithm requires less computation compared with the other two algorithms, as only a summation over an $N/2$-dimensional vector is needed for each enumerated time delay, while mask-weighted GCC-PHAT requires a summation over all the unit pairs and mask-weighed steered-response SNR needs a series of matrix multiplications.

Previous studies [8], [42], [43] have computed time delays from estimated steering vectors at each frequency band or each T-F unit pair. They divide the estimated phase difference by the angular frequency to get the time delay, assuming that the microphones are placed sufficiently close and no phase wrapping occurs. However, using closely spaced microphones would make the time delay too small to be accurately estimated and also make location triangulation harder. When phase wrapping is present, multiple time delays could give exactly the same phase difference at a specific frequency band. Our method addresses this ambiguity via enumerating all the time delays and checking the similarity measure in Eq. (18) of each time delay. This method is sensible because a time delay deterministically corresponds to a phase difference. Another difference is that we use DNN based T-F masking for steering vector computation. In contrast, previous studies use spatial clustering [42] or empirical rules [43].

Our proposed algorithm differs from the classic MUSIC algorithm [4] and its recent extension in [22] where a recurrent neural network with uni-directional long short-term memory (LSTM) is used to estimate the ideal binary mask and the estimated mask is then utilized to weight spatial covariance matrix estimation for MUSIC. Whereas these studies find the target direction with its hypothesized steering vector orthogonal to the noise subspace, the proposed algorithm directly searches for a direction that is closely matched to target steering vectors between each pair of microphones at all frequencies. The steering vector in our study is robustly estimated using supervised T-F masking. Similar to GCC-PHAT, our algorithm implicitly equalizes the contribution of each frequency as all frequencies contain information for the DOA estimation of broadband speech signals. In contrast, the pseudospectram at each frequency in the broadband MUSIC algorithm used in [22] is unbounded, and some frequencies could dominate the summation of the pseudospectrams.

D. Deep Learning Based Mask Estimation

Clearly, the estimated mask of each microphone signal, M_p, plays an essential role in the proposed algorithms. Deep learning based T-F masking has advanced monaural speech separation and enhancement performance by large margins [19]. Many DNNs have been applied to time-frequency masking. Among them, RNNs with bi-directional LSTM (BLSTM) have shown consistently better performance over feed-forward neural networks, convolutional neural networks, simple RNNs [44], and RNNs with uni-directional LSTM [45], [46], due to their better modeling of contextual information. In this study, we train an RNN with BLSTM to estimate the IRM (see Section III for more details of BLSTM training). When computing the IRM of a noisy and reverberant utterance, we consider the direct sound as the target signal and the remaining components as interference, as the direct sound contains phase information for DOA

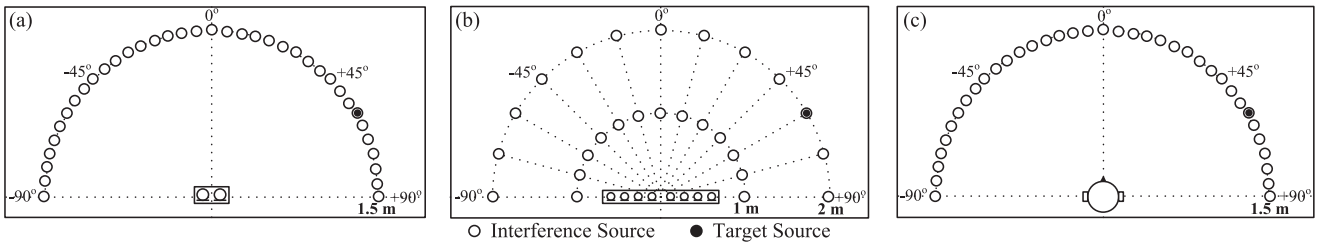

Fig. 2. Illustration of the (a) two-microphone setup, (b) eight-microphone setup, and (c) binaural setup.

estimation.

$$\mathrm{IRM}_p\left(t,f\right) = \sqrt{\frac{\left|c_p\left(f\right)s\left(t,f\right)\right|^2}{\left|c_p\left(f\right)s\left(t,f\right)\right|^2+\left|h_p\left(t,f\right)+n_p\left(t,f\right)\right|^2}} \tag{21}$$

See Eq. (1) for relevant notations in the above equation. In single-channel speech enhancement, the estimated real-valued mask is element-wise multiplied with the STFT coefficients of unprocessed noisy speech to obtain enhanced speech [33]. In this study, we use an estimated IRM to weight T-F units for DOA estimation. Our study uses log power spectrogram features for mask estimation.

The IRM is *ideal* for speech enhancement only when the mixture phase is the same as the clean phase at each T-F unit. The phase-sensitive mask (PSM) [34], [35] takes the phase difference into consideration by scaling down the ideal mask when the mixture phase is different from the clean phase using a cosine operation. In a way, it represents the best mask if a real-valued mask is multiplied with the STFT coefficients of unprocessed noisy speech for enhancement [34], [47]. We define a form of the phase-sensitive mask in the following way:

$$\begin{aligned}\mathrm{PSM}_p\ \left(t,f\right) = \max\{0, \mathrm{IRM}_p\left(t,f\right)\ \cos\left(\angle y_p\left(t,f\right)\right.\\ \left.-\angle\left(c_p\left(f\right)s\left(t,f\right)\right)\right)\}\end{aligned} \tag{22}$$

The inclusion of phase in an ideal mask seems particularly suited for our task as phase is key for localization and we need to identify T-F units with cleaner phase for this task. The cosine term serves to reduce the contributions of contaminated T-F units for localization. Note the difference between the PSM defined in Eq. (22) and the definition in [34].

III. Experimental Setup

The proposed localization algorithms are evaluated in reverberant environments with strong diffuse babble noise. Our neural network is trained only on simulated room impulse responses (RIR) using just single-channel information for mask estimation, and directly tested on three unseen sets of RIRs for DOA estimation using microphone arrays with various numbers of microphones arranged in diverse ways. An illustration of the test setup is shown in Fig. 2. The first test set includes a relatively matched set of simulated two-microphone RIRs, the second set consists of real RIRs measured on an eight-microphone array, and the third set contains real binaural RIRs (BRIR) measured on a dummy head.

The RIRs used in the training and validation data are simulated using an RIR generator[1], which is based on the classic image method. An illustration of this setup is shown in Fig. 2(a). For the training and validation set, we place 36 different interfering speakers at the 36 directions uniformly spaced between -87.5° and 87.5° in steps of 5°, i.e., one competing speaker in each direction, resulting in a 36-talker diffuse babble noise. The target speaker is randomly placed at one of the 36 directions. For the testing data, we put 37 different interference speakers at the 37 directions spanning from -90° to 90° in steps of 5° (one competing speaker in each direction), and the target speaker randomly at one of the 37 directions. This way, the test RIRs are different from the RIRs used for training and validation. The distance between each speaker and the array center is 1.5 m (see Fig. 2(a)). The room size is fixed at 8x8x3 m, and the two microphones are placed around the center of the room. The spacing between the two microphones is 0.2 m and the microphone heights are both set to 1.5 m. The reverberation time (T60) of each mixture is randomly selected from 0.0 s to 1.0 s in steps of 0.1 s. Target speech comes from the IEEE corpus with 720 sentences uttered by a female speaker [48]. We split the utterances into sets of 500, 100 and 120 (in the same order as listed in the IEEE corpus) to generate training, validation and test data. To create the diffuse babble noise for each mixture, we randomly pick 37 (or 36) speakers from the 462 speakers in the TIMIT training set and concatenate all the utterances of each speaker, and then place them at all 37 (or 36) directions, with a randomly chosen speech segment of each speaker per direction. Note that we use the first half of the concatenated utterance of each speaker to generate the training and validation diffuse babble noise, and the second half to generate the test diffuse noise. There are in total 50,000, 1,000, and 3,000 two-channel mixtures in the training, validation and test set, respectively. The average duration of the mixtures is 2.4 s. The input SNR computed from reverberant speech and reverberant noise is fixed at -6 dB. Note that if the direct sound is considered as target speech and the remaining signal as noise, as is done in Eq. (21) and (22), the SNR will vary a lot and be much lower than -6 dB, depending on the direct-to-reverberant ratio (DRR)[2] of the RIRs. We therefore fix the SNR between the reverberant speech and reverberant noise at -6 dB and systematically vary the RIRs to change the SNR between the direct sound signal and the remaining components.

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

[2] For a simulated RIR, the corresponding simulated anechoic RIR is considered as the RIR of direct sound. For a real RIR, we first find the sample ℓ with the largest absolute value. Then the RIR of direct sound is considered as the first $\ell + 0.0025 f_s$ samples for DRR and IRM/PSM computation.

We train our BLSTM using all the single-channel signals (50,000 $\times$ 2 in total) in the training data. The log power spectrogram is used as the input features for mask estimation. Global mean-variance normalization is performed on the input features. The BLSTM consists of two hidden layers each with 600 units in each direction. Sigmoidal units are utilized in the output layer, as the IRM and PSM are bounded between zero and one. During training, the Adam algorithm is utilized to minimize the mean squared error for a maximum of 100 epochs, starting with an initial learning rate of 0.001, which is scaled by half if the error on the validation set is not reduced after three epochs. The frame length is 32 ms, the frame shift is 8 ms, and the sampling rate is 16 kHz. A 512-point FFT (fast Fourier transform) is performed to extract 257-dimensional log spectrogram feature at each frame. The input and output dimension are thus both 257. The sequence length for BLSTM training and testing is just the utteance length.

The proposed algorithms are also evaluated on the Multi-Channel Impulse Responses Database [49][3] measured at Bar-Ilan University using a set of eight-microphone linear arrays. We use the microphone array with 8 cm spacing between the two center microphones, and 4 cm spacing between the other adjacent microphones in our experiments, i.e., 4-4-4-8-4-4-4. The setup is depicted in Fig. 2(b). The RIRs are measured in a room with the size 6 $\times$ 6 $\times$ 2.4 m in steps of 15° from -90° to 90°, at a distance of 1.0 and 2.0 m to the array center, and at three reverberation time (0.16, 0.36 and 0.61 s). Similar to the two-microphone setup, the IEEE and TIMIT utterances are utilized to generate 3,000 eight-channel test utterances for each of the two distances. We put one interference speaker at each of the 26 locations, resulting in a 26-talker diffuse babble noise. For each of the two distances, the target speaker is placed at one of the 11 interior locations on the hemi-circle (to avoid endfire directions). Note that the RIRs, number of microphones, source-to-array distance, and microphone geometry in this dataset are all unseen during training. In addition, the diffuse babble noise is generated using different locations and different number of interfering speakers. The trained BLSTM is directly tested on the generated test utterances using randomly selected sets of microphones to demonstrate the versatility of our approach to arrays with varying numbers of microphones arranged in diverse geometries.

We also evaluate our algorithm on a binaural setup illustrated in Fig. 2(c). The real BRIRs[4] captured using a Cortex head and torso simulator (HATS dummy head) in four real rooms with different sizes and T60s at the University of Surrey are utilized to generate the test utterances. The dummy head is placed at various heights between 1.7 m and 2.0 m in each room, and the source to array distance is 1.5 m. The real BRIRs are measured using 37 directions ranging from -90° to 90° in steps of 5°. The IEEE and TIMIT utterances are utilized to generate 3,000 binaural test utterances in the same way as in the two-microphone setup. The only difference from the two-microphone setup illustrated in Fig. 2(a) is that now real BRIRs rather than simulated two-channel RIRs are used to generate test utterances. Note that we directly apply the trained BLSTM on this new binaural test set for DOA estimation, although the BLSTM is not trained specifically on any binaural data and the binaural setup is completely unseen during training.

For setup (a) and (b), the location or direction of interest k is enumerated from -90° to 90° in steps of 1° on the hemi-circle. The hypothesized time delay between microphone p and q for location or direction k, $\tau_{p,q}(k)$, is computed as $(d_{kq} - d_{kp})/c_s$, where c_s is 343 m/s in the air. Note that setup (b) uses real RIRs measured by a given microphone array, so the distance between each candidate location and each microphone, and microphone configurations are all subject to inaccuracies. In addition, the assumed sound speed may not equal the actual sound speed. These factors complicate accurate localization. For setup (c), the hypothesized time delay cannot be obtained from the distance difference due to the shadowing of head and torso. $\tau_{1,2}(k)$ is instead enumerated from -15 to 15 samples in steps of 0.1 sample. The estimated time delay is then mapped to the azimuth giving the closest time delay. This mapping is obtained from the group delay of the measured BRIRs of the HATS dummy head in the anechoic condition, as is done in [16].

Note that we assume that the target speaker is fixed within each utterance (average length is 2.4 s), and compute a single DOA estimate per utterance. For setup (a) and (c), which use 5$^\text{o}$ step size for the candidate directions, we measure localization performance using gross accuracy, which considers a prediction correct if it is within 5$^\text{o}$ (inclusive) of the true target direction. For the Multi-Channel Impulse Response Database with a coarser spatial resolution, we consider a prediction correct if it is within 7.5° of the true direction. Gross accuracy is given as percent correct over all test utterances.

In Eq. (6), (14) and (19), Ω contains all the microphone pairs of an array for the summation.

IV. Evaluation Results

Table I presents localization gross accuracy results for two-microphone setup (a), together with the DRR at each T60 and the oracle performance marked in grey. We report DRR together with T60 as it is an important factor for the performance of sound localization in reverberant environments. The rows of eIRM and ePSM in the table mean that estimated IRM and estimated PSM are used for DOA estimation, respectively. All the three proposed algorithms lead to large improvements over classic GCC-PHAT and MUSIC algorithms (on average 72.0%, 86.7% and 75.1% using ePSM vs. 21.6% and 25.2%). PSM estimation yields consistently better performance than IRM estimation for all the algorithms; similar trends are observed from later results in Tables II–IV. As is reported in Table I, frequency weighting based on estimated masks, i.e., using Eq. (16) and (20), leads to consistent improvements (more than 5 percentage points on average). Among the three proposed algorithms, mask-weighted steered-response SNR performs the best, especially when reverberation time is high and DRR is low. For all the three proposed algorithms, using the PSM or IRM results in close to 100% gross accuracy, even when reverberation time is as high as 1.0 s, the DRR is as low as -8.0 dB, and the SNR between reverberant speech and reverberant noise is as

[3]Available at http://www.eng.biu.ac.il/gannot/downloads/

[4]Available at https://github.com/IoSR-Surrey/RealRoomBRIRs

TABLE I
DOA ESTIMATION PERFORMANCE (% GROSS ACCURACY) OF DIFFERENT METHODS IN TWO-MICROPHONE SETUP

Method	Frequency Weighting	Mask	T60(s)/DRR(dB)										AVG
			0.0/Inf	0.2/3.8	0.3/-0.4	0.4/-2.5	0.5/-4.0	0.6/-5.1	0.7/-6.0	0.8/-6.8	0.9/-7.4	1.0/-8.0	
GCC-PHAT	-	-	33.7	35.6	30.1	26.1	16.7	15.6	19.5	14.3	15.2	8.9	21.6
MUSIC	-	-	35.1	41.6	33.9	26.7	20.6	20.5	23.6	16.7	19.3	13.9	25.2
Mask-weighted GCC-PHAT	-	eIRM	94.3	95.7	87.0	80.1	74.6	64.0	53.4	49.0	47.2	38.6	68.3
	-	IRM	99.3	99.7	98.7	96.1	96.9	97.1	96.8	94.9	96.2	95.7	97.1
	-	ePSM	96.4	95.4	88.3	82.7	80.1	69.2	59.1	53.7	51.0	44.6	72.0
	-	PSM	100.0	100.0	100.0	100.0	100.0	99.7	99.7	99.3	100.0	99.3	99.8
Mask-weighted Steered-response SNR	Eq. (15)	eIRM	94.6	93.7	84.8	78.5	80.1	80.2	68.1	59.5	59.7	57.8	75.7
	Eq. (16)	eIRM	95.0	95.0	87.7	84.0	85.7	87.7	75.7	69.7	66.6	64.7	81.2
	Eq. (16)	IRM	100.0	99.7	99.1	99.3	99.3	99.4	99.4	99.3	99.3	99.3	99.4
	Eq. (15)	ePSM	94.6	95.4	87.0	82.7	87.1	84.7	75.1	65.6	66.6	62.0	80.1
	Eq. (16)	ePSM	96.1	96.4	91.1	89.6	91.3	89.0	84.0	76.9	76.9	75.9	86.7
	Eq. (16)	PSM	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.7	100.0
DOA Estimation from Steering Vectors	Eq. (19)	eIRM	89.6	92.4	84.2	73.3	70.4	64.6	55.6	51.4	50.0	40.6	67.2
	Eq. (20)	eIRM	93.5	95.7	86.4	80.8	76.7	69.2	61.0	58.8	55.2	47.2	72.4
	Eq. (20)	IRM	98.9	99.7	99.1	97.1	97.2	96.8	96.2	94.2	95.9	96.4	97.1
	Eq. (19)	ePSM	90.7	92.4	84.5	76.5	72.5	67.9	60.1	51.4	50.7	43.9	69.0
	Eq. (20)	ePSM	96.1	97.0	88.3	82.7	80.8	70.5	66.1	58.8	57.2	54.1	75.1
	Eq. (20)	PSM	99.6	100.0	100.0	100.0	100.0	99.7	99.7	99.3	99.7	99.3	99.7

TABLE II
DOA ESTIMATION PERFORMANCE (%GROSS ACCURACY) OF DIFFERENT METHODS IN MULTI-MICROPHONE SETUP BY RANDOMLY SELECTING TWO MICROPHONES FOR EACH TEST UTTERANCE

Method	Mask	Distance	T60(s)/DRR(dB)			AVG	Distance	T60(s)/DRR(dB)			AVG
			0.16/10.5	0.36/7.4	0.61/4.7			0.16/6.3	0.36/1.6	0.61/-1.3	
GCC-PHAT	-	1 m	37.9	38.5	31.7	36.1	2m	31.7	29.8	22.8	28.1
MUSIC	-		34.6	35.8	31.7	34.1		30.0	23.9	21.1	25.0
Mask-weighted GCC-PHAT	eIRM		84.0	83.0	84.3	83.7		82.1	74.4	67.5	74.6
	IRM		92.7	91.6	93.0	92.4		92.8	93.0	91.7	92.5
	ePSM		85.6	85.9	83.0	84.9		85.2	78.3	70.6	78.1
	PSM		94.0	94.0	92.5	93.5		93.3	93.2	92.4	93.0
Mask-weighted Steered-response SNR	eIRM		84.0	84.2	82.5	83.6		81.5	69.3	66.6	72.4
	IRM		93.2	92.9	92.7	92.9		93.1	92.2	92.7	92.6
	ePSM		86.7	86.5	86.4	86.5		85.0	77.1	72.4	78.2
	PSM		92.8	93.8	92.5	93.1		95.2	92.6	91.8	93.2
DOA Estimation from Steering Vectors	eIRM		80.4	80.6	81.6	80.8		79.5	67.9	65.3	70.9
	IRM		92.3	90.2	92.8	91.7		92.8	92.6	91.1	92.2
	ePSM		83.6	83.4	81.3	82.8		81.9	73.8	68.1	74.6
	PSM		93.8	93.9	92.2	93.3		92.9	92.9	92.4	92.8

TABLE III
DOA ESTIMATION PERFORMANCE (%GROSS ACCURACY, AVERAGED OVER ALL REVERBERATION TIMES) OF DIFFERENT METHODS AT 2 M DISTANCE IN MULTI-MICROPHONE SETUP BY RANDOMLY SELECTING DIFFERENT NUMBERS OF MICROPHONES FOR EACH TEST UTTERANCE.

Method	Mask	# microphones						
		2	3	4	5	6	7	8
GCC-PHAT	-	28.1	36.1	38.9	41.8	41.5	41.4	42.8
MUSIC	-	25.0	30.4	31.3	32.2	32.8	32.7	32.8
Mask-weighted GCC-PHAT	eIRM	74.6	89.3	93.8	94.6	95.1	96.0	96.1
	IRM	92.5	98.2	99.6	100.0	100.0	100.0	100.0
	ePSM	78.1	90.3	93.7	95.5	95.9	96.2	96.2
	PSM	93.0	98.7	99.7	100.0	100.0	100.0	100.0
Mask-weighted Steered-response SNR	eIRM	72.4	85.8	90.1	92.1	92.9	93.4	93.5
	IRM	92.6	98.7	99.6	100.0	100.0	100.0	100.0
	ePSM	78.2	90.0	93.5	94.7	95.6	95.8	95.8
	PSM	93.2	98.9	99.8	100.0	100.0	100.0	100.0
DOA Estimation from Steering Vectors	eIRM	70.9	85.6	89.8	91.3	92.2	92.4	92.6
	IRM	92.2	98.3	99.6	100.0	100.0	100.0	100.0
	ePSM	74.6	88.9	92.6	94.4	94.8	95.1	95.1
	PSM	92.8	98.7	99.7	100.0	100.0	100.0	100.0

TABLE IV
DOA ESTIMATION PERFORMANCE (% GROSS ACCURACY) OF DIFFERENT METHODS IN BINAURAL SETUP

Method	Mask	Room - T60(s)/DRR(dB)					AVG
		Anechoic 0.0/Inf	A 0.32/7.2	B 0.47/7.0	C 0.68/10.9	D 0.89/7.3	
GCC-PHAT	-	56.7	28.7	36.6	33.4	25.3	36.0
MUSIC	-	56.4	26.0	36.1	28.0	26.1	34.3
Mask-weighted GCC-PHAT	eIRM	96.6	94.7	94.8	95.1	91.2	94.5
	IRM	100.0	99.4	99.8	99.3	100.0	99.7
	ePSM	97.4	95.3	96.6	95.6	94.3	95.8
	PSM	100.0	99.5	100.0	99.3	100.0	99.8
Mask-weighted Steered-response SNR	eIRM	96.6	88.6	89.1	87.6	85.8	89.5
	IRM	99.7	99.5	99.5	99.2	99.8	99.5
	ePSM	97.4	93.6	93.8	89.3	90.4	92.9
	PSM	100.0	100.0	99.7	99.8	100.0	99.9
DOA Estimation from Steering Vectors	eIRM	97.6	91.3	91.3	86.0	85.2	90.2
	IRM	100.0	99.4	99.8	99.2	99.8	99.6
	ePSM	97.6	95.3	93.9	89.6	89.9	93.3
	PSM	100.0	99.5	99.8	99.0	100.0	99.7

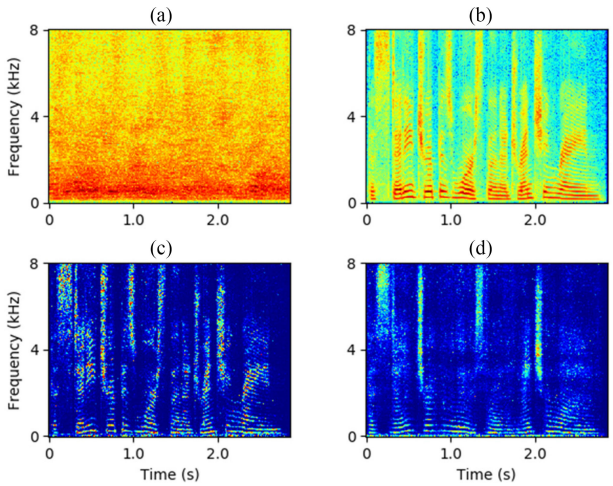

Fig. 3. Illustration of an estimated IRM for a mixture with babble noise in the two-microphone setup (SNR = −6 dB and T60 = 0.9 s). (a) Mixture log power spectrogram; (b) clean log power spectrogram; (c) IRM; (d) estimated IRM.

low as −6 dB. These oracle results demonstrate the effectiveness of T-F masking: the PSM and IRM can be considered as strong training targets for robust speaker localization, just like for speech separation and enhancement [50], [33]. In addition, better estimated masks in the future will likely produce better localization results.

For the mask-weighted GCC-PHAT algorithm, we have also evaluated the average of estimated mask instead of the product in Eq. (5), motivated by [21]. We find that the product rule produces significantly better localization than the average, 68.3% vs. 55.3% using eIRM and 72.0% vs. 61.6% using ePSM. We should note that the average mask is not exactly what is used in [21] and there are many differences between our system and [21], as discussed in Section II-A. These differences complicate a direct comparison. Another way is to compare the relative improvement over a baseline where no masking is performed. It appears that our overall system obtains larger improvements.

Fig. 3 illustrates IRM estimation for a very noisy and reverberant mixture. As can be observed by comparing the IRM in Fig. 3(c) and the estimated IRM in Fig. 3(d), the estimated mask well resembles the ideal mask in this case, indicating the effectiveness of BLSTM based mask estimation. Upon a closer examination, we observe that the IRM is more accurately estimated at speech onsets and lower frequencies, likely because the direct speech energy is relatively stronger in these T-F regions.

Table II presents the accuracy of DOA estimation in setup (b), which uses measured real RIRs. For each utterance, we randomly choose two microphones from the eight microphones for testing. Note that the microphone distances can vary from 4 cm at minimum to 28 cm at maximum for the test utterances. As the DNN in our algorithms only utilizes single-channel information, our approach can still apply even as geometry varies substantially. As can be seen, the proposed algorithms using PSM lead to large improvements over GCC-PHAT and MUSIC, 84.9%, 86.5% and 82.8% vs. 36.1% and 34.1% for 1 m distance, and 78.1%, 78.2% and 74.6% vs. 28.1% and 25.0% for 2 m distance. In this setup, the three proposed algorithms perform similarly, with the mask-weighted steered-response SNR performing slightly better. Clearly, the performance is better when the source to array distance is 1 m than 2 m. Using the IRM or the PSM does not reach 100% accuracy in this setup, likely because the aperture size can be as small as 4 cm, posing a fundamental challenge for accurate localization of a distant speaker.

In Table III, we show that our algorithms can be directly extended to multi-channel cases. This is done by combining different microphone pairs as in the classic SRP-PHAT algorithm. For each utterance, we randomly select a number of microphones for testing. As can be seen from the results, using more microphones leads to better performance for all the algorithms. A significant improvement occurs going from two to three microphones, likely because three microphone pairs become available for localization in a three-sensor array versus one pair in a two-sensor array. The performance starts to plateau after five microphones. Among the proposed algorithms, the mask-weighted GCC-PHAT algorithm performs slightly better than the other two when more microphones become available.

Table IV reports the results on binaural setup (c). Although the neural network trained for mask estimation has not seen binaural signals and binaural geometry, directly applying it to binaural speaker localization results in substantial gains over the GCC-PHAT and MUSIC algorithms. Notably, the mask-weighted steered-response SNR algorithm is slightly worse than the other two (92.9% vs. 95.8% and 93.3% using ePSM). The reason, we think, is that the energy levels at the two channels cannot be treated as equal as is done in Eq. (10), as head shadow effects occur in the binaural setup. For the microphone array setup (a) and (b), assuming equal energy levels is reasonable as there is no blockage from sound sources to an array. Also the localization performance in this binaural setup appears much higher than the two-microphone setup, likely because the DRR is much higher.

In the above localization evaluations, the IEEE utterances of the same speaker are used in both training and testing. How sensitive is our approach to a training speaker? To get an idea, we evaluate the performance of the already-trained model without any change on a new IEEE female speaker in the two-microphone setup. In this evaluation, the last 120 sentences

TABLE V
DOA ESTIMATION PERFORMANCE (% GROSS ACCURACY) OF DIFFERENT METHODS IN TWO-MICROPHONE SETUP ON A NEW IEEE FEMALE SPEAKER

Method	Frequency Weighting	Mask	T60(s)/DRR(dB)										AVG
			0.0/Inf	0.2/3.8	0.3/-0.4	0.4/-2.5	0.5/-4.0	0.6/-5.1	0.7/-6.0	0.8/-6.8	0.9/-7.4	1.0/-8.0	
GCC-PHAT	-	-	49.8	51.5	45.8	33.1	32.5	23.7	17.5	16.9	12.2	15.8	29.8
MUSIC	-	-	66.8	70.9	56.3	44.5	46.9	34.6	25.3	27.9	17.9	28.1	41.8
Mask-weighted GCC-PHAT	-	eIRM	91.9	90.9	86.4	81.8	77.3	59.2	55.9	51.7	40.1	33.8	66.8
	-	IRM	100.0	100.0	99.7	98.8	100.0	98.1	98.3	96.9	96.8	96.8	98.5
	-	ePSM	93.7	94.8	88.8	86.0	81.1	66.0	66.7	58.6	53.0	53.0	74.2
	-	PSM	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Mask-weighted Steered-response SNR	Eq. (15)	eIRM	87.5	84.5	72.9	67.2	71.3	62.3	57.9	57.6	48.4	50.2	65.8
	Eq. (16)	eIRM	91.5	90.3	79.0	76.4	81.8	69.5	64.3	67.2	55.6	59.6	73.4
	Eq. (16)	IRM	99.6	100.0	100.0	100.0	99.7	99.1	99.7	99.3	99.6	99.7	99.7
	Eq. (15)	ePSM	89.3	93.9	87.8	84.8	81.1	71.7	73.7	68.3	62.0	65.6	77.8
	Eq. (16)	ePSM	97.0	97.4	89.8	89.6	87.4	76.6	80.8	75.9	71.7	74.1	84.0
	Eq. (16)	PSM	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.7	100.0	100.0	100.0
DOA Estimation from Steering Vectors	Eq. (19)	eIRM	91.9	90.3	86.1	78.5	75.9	62.3	62.0	57.2	44.1	49.5	69.7
	Eq. (20)	eIRM	93.0	93.9	88.8	85.4	80.8	66.0	66.0	59.7	53.8	54.6	74.2
	Eq. (20)	IRM	99.6	100.0	99.7	98.5	99.7	97.5	98.0	97.2	97.5	96.2	98.4
	Eq. (19)	ePSM	95.6	93.5	88.1	81.8	80.4	64.8	67.0	61.0	49.1	54.9	73.6
	Eq. (20)	ePSM	95.2	96.4	91.9	87.8	84.3	72.6	71.0	66.6	54.8	62.5	78.3
	Eq. (20)	PSM	100.0	100.0	100.0	100.0	100.0	100.0	99.7	99.7	100.0	100.0	99.9

uttered by the new speaker are used to generate 3,000 test utterances in the same way as in the evaluation of Table I. The results on the new speaker are given in Table V. As can be observed from this table, the relative improvements of the proposed algorithms over the baseline GCC-PHAT and MUSIC are, as expected, not as large as those for the training speaker in Table I (the estimated masks are less accurate in this case), but they are still substantial.

V. CONCLUDING REMARKS

We have investigated a new approach to robust speaker localization that is guided by T-F masking. Benefiting from monaural masking based on deep learning, our approach dramatically improves the robustness of conventional cross-correlation-, beamforming- and subspace-based approaches for speaker localization in noisy and reverberant environments. We have found that balancing the contribution of each frequency is important for the DOA estimation of broadband speech signals. Although the neural network is trained using single-channel information, our study shows that it is versatile in its application to arrays with various numbers of microphones and diverse geometries. This property should be useful for cloud-based services, where client setup may vary significantly in terms of microphone configuration.

Our approach has a number of limitations that need to be dealt with in future work. Our system performs localization at the utterance level, and as a result it cannot be applied to localizing moving sound sources. The current study does not locate multiple speakers. BLSTM is inherently non-causal and cannot be used for online applications where uni-directional LSTM is more appropriate. Although our evaluations have employed recorded RIRs, we have not used recorded speech signals. In addition, one noise type and one SNR are considered in our current study, although SNR generalization is not expected to be difficult [19] and noise generalization can be addressed via large-scale training [51], [52].

Before closing, we emphasize that the proposed approach achieves robust speaker localization as guided by T-F masking. Our experiments find that even for severely corrupted utterances, ratio masking in the proposed algorithms leads to accurate localization. Our study suggests that ideal ratio masks can serve as strong training targets for robust speaker localization. Clearly, the major factor limiting the localization performance is the quality of estimated masks. Nonetheless, the proposed T-F masking guided approach promises further localization improvements as robust speaker localization can directly benefit from the rapid development of deep learning based time-frequency masking. Through training, masking guidance plays the dual role of specifying the target source and attenuating sounds interfering with localization. T-F masking affords a view of the signal to be localized, as opposed to traditional localization that blindly relies on signal energy.

ACKNOWLEDGMENT

The authors would like to thank Dr. S. Charkrabarty for helpful discussions on microphone array processing, and the anonymous reviewers for their constructive comments.

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 Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 692–730, Apr. 2017.

[2] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, pp. 320–327, Aug. 1976.

[3] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*. Berlin, Germany: Springer, 2001, pp. 157–180.

[4] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.

[5] J.-M. Valin, F. Michaud, J. Rouat, and D. Letourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2003, vol. 2, pp. 1228–1233.

[6] Y. Rui and D. Florencio, "Time delay estimation in the presence of correlated noise and reverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, pp. 133–136.
[7] H.-G. Kang, M. Graczyk, and J. Skoglund, "On pre-filtering strategies for the GCC-PHAT algorithm," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2016, pp. 1–5.
[8] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, 2003.
[9] J. M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robot. Auton. Syst.*, vol. 55, no. 3, pp. 216–228, 2007.
[10] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Process.*, vol. 92, no. 8, pp. 1950–1960, 2012.
[11] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA, USA: MIT Press, 1990.
[12] W. M. Hartmann, "How we localize sound," *Phys. Today*, vol. 52, no. 11, pp. 24–29, 1999.
[13] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2006.
[14] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *J. Acoust. Soc. Amer.*, vol. 123, pp. 2136–2147, 2008.
[15] W. Zhang and B. D. Rao, "A two microphone-based approach for source localization of multiple speech sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, pp. 1913–1928, Nov. 2010.
[16] J. Woodruff and D. L. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1503–1512, Jul. 2012.
[17] F. Grondin and F. Michaud, "Time difference of arrival estimation based on binary frequency mask for sound source localization on mobile robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 6149–6154.
[18] W. Zheng, Y. Zou, and C. Ritz, "Spectral mask estimation using deep neural networks for inter-sensor data ratio model based robust DOA estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 325–329.
[19] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.
[20] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.
[21] P. Pertila and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 6125–6129.
[22] 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 *Proc. Interspeech*, 2017, pp. 1894–1898.
[23] Z.-Q. Wang, X. Zhang, and D. L. Wang, "Robust TDOA estimation based on time-frequency masking and deep neural networks," in *Proc. Interspeech*, 2018, pp. 322–326.
[24] G. C. Carter, A. H. Nuttall, and P. G. Cable, "The smoothed coherence transform," *Proc. IEEE*, vol. 61, no. 10, pp. 1497–1498, Oct. 1973.
[25] C. Zhang, D. Florêncio, and Z. Zhang, "Why does PHAT work well in lownoise, reverberative environments?," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 2565–2568.
[26] N. Ma, G. Brown, and T. May, "Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions," in *Proc. Interspeech*, 2015, pp. 160–164.
[27] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 2814–2818.
[28] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 136–140.
[29] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 2386–2390.
[30] N. Ma, T. May, and G. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 12, pp. 2444–2453, Dec. 2017.
[31] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 444–451.
[32] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 276–280.
[33] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.
[34] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 708–712.
[35] Y. Wang and D. L. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4390–4394.
[36] V. Krishnaveni, T. Kesavamurthy, and B. Aparna, "Beamforming for direction-of-arrival (DOA) estimation - A survey," *Int. J. Comput. Appl.*, vol. 61, no. 11, pp. 975–8887, 2013.
[37] M. Zohourian, G. Enzner, and R. Martin, "Binaural speaker localization integrated in an adaptive beamformer for hearing aids," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 515–528, Mar. 2018.
[38] T. Yoshioka *et al.*, "The NTT CHiME-3 System: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 436–443.
[39] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, 2015, pp. 504–511.
[40] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Comput. Speech Lang.*, vol. 46, pp. 535–557, 2017.
[41] X. Anguera and C. Wooters, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2011–2022, Sep. 2007.
[42] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, pp. 33–36.
[43] N. T. N. Tho, S. Zhao, and D. L. Jones, "Robust DOA estimation of multiple speech sources," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 2287–2291.
[44] Z.-Q. Wang and D. L. Wang, "Recurrent deep stacking networks for supervised speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 71–75.
[45] F. Weninger *et al.*, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. Int. Conf. Latent Var. Anal. Signal Separation*, 2015, pp. 91–99.
[46] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[47] D. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," in *Proc. IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2016, pp. 483–492.
[48] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AE-17, no. 3, pp. 225–246, Sep. 1969.
[49] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2014, pp. 313–317.
[50] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. New York, NY, USA: Springer, 2005, pp. 181–197.
[51] J. Chen, Y. Wang, S. E. Yoho, D. L. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Amer.*, vol. 139, no. 5, pp. 2604–2612, 2016.
[52] J. Chen and D. L. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *J. Acoust. Soc. Amer.*, vol. 141, pp. 4705–4714, 2017.

Authors' photographs and biographies not available at the time of publication.