ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095551

GCC-SPEAKER: TARGET SPEAKER LOCALIZATION WITH OPTIMAL SPEAKER-DEPENDENT WEIGHTING IN MULTI-SPEAKER SCENARIOS

Guanjun Li[1], *Wei Xue*[2], *Wenju Liu*[1], *Jiangyan Yi*[1], *Jianhua Tao*[3]

[1]NLPR, Institute of Automation, Chinese Academy of Sciences, China
[2]Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
[3]Department of Automation, Tsinghua University

ABSTRACT

Existing noise-robust and reverberant-robust localization algorithms fail to localize the target speaker when interfering speakers are present. In this paper, we address the problem of localizing only the target speaker in multi-speaker scenarios and propose a target speaker localization algorithm, called GCC-speaker. Specifically, we modify the weighting of the generalized cross-correlation with phase transform (GCC-PHAT) algorithm and propose an optimal speaker-dependent weighting based on a novel localization-related loss function and data-driven training. The speaker-dependent weighting is responsible for guiding the GCC algorithm to obtain the optimal target speaker localization results. As for the loss function, we constrain the estimated GCC angular spectrum and the estimated direction of arrival (DOA) to be close to their ground truth values, respectively. The experimental results show the superiority of GCC-speaker compared to the existing target speaker localization algorithms for different signal-to-interference ratios, reverberation times and array geometries.

***Index Terms*—** Target speaker localization, speaker-dependent weighting, generalized cross-correlation

1. INTRODUCTION

Speaker localization aims to estimate the direction of arrival (DOA) of a speaker from microphone signals, and it is of importance for multi-channel speech enhancement and recognition [1]. The generalized cross-correlation with phase transform (GCC-PHAT) [2] and multiple signal classification (MUSIC) [3] are the two most popular speaker localization algorithms, among which GCC-PHAT has become a mainstream mainly because of its computational efficiency and good tracking capability [4]. However, the performance of GCC-PHAT is still unsatisfactory in adverse scenarios, which may include noise, reverberation and multiple speakers [5].

Although many approaches have been developed to improve the robustness of GCC-PHAT against noise or reverberation [6, 7, 8, 9, 10], the multi-speaker issue (e.g., in the cocktail party) still remains challenging, because interfering speakers are non-negligible factors and prevent the GCC-PHAT from effectively localizing the speaker of interested. Localizing the target speaker in multi-speaker cases is called *target speaker localization* [11, 12, 13] and it differs with the *multi-speaker localization* [14, 15, 16] which estimates the DOAs of all speakers. Although the target speaker may be localized in the multi-speaker localization settings with an additional post-identification process, the multi-speaker localization itself always relies on counting source numbers and can be erroneous. Therefore, directly identifying the target speaker's DOA is more practical and can be widely used in applications such as teleconferencing and smart personal devices which concern only one target speaker in probably overlapping speeches. In this paper, we will address the target speaker localization problem.

There are several works [11, 12, 13] for target speaker localization. These methods generally first estimate a speaker-dependent mask and then apply it to the existing localization algorithms. In [11] and [12], the speech enhancement loss and the keyword cues are used to obtain the speaker-dependent mask, respectively. In [13], an external microphone close to the target speaker is adopted to statistically derive the speaker-dependent mask combining with the original microphone array signals. Although many strategies can be used to derive the mask, the designing of these mask estimation is rather heuristic, and it is hard to judge whether these masks are optimal for target speaker localization. As a consequence, these methods usually yield sub-optimal results.

In this paper, we tightly integrate the speaker-dependent mask estimator and the computationally efficient GCC-PHAT algorithm to achieve end-to-end target speaker localization. The proposed algorithm is called GCC-speaker. Compared with previous target speaker localization works [11, 12, 13], the mask estimator is directly optimized with a localization-related loss function. Compared with previous robust GCC-PHAT algorithm [6, 7, 8, 9, 10], we propose an optimal speaker-dependent weighting for GCC in multi-speaker scenarios. More specifically, we use the well-known SpeakerBeam [17] structure as the mask estimator to output the speaker-dependent mask. Next, we combine the speaker-dependent mask with the weighting function of GCC-PHAT to propose an optimal speaker-dependent weighting based on a novel localization-related loss function. The optimal speaker-dependent weighting is then used by the GCC algorithm to obtain the optimal DOA results. As for the loss function, we adopt a multi-task learning (MTL) strategy to make the estimated GCC angular spectrum and the estimated DOA close to their ground truth values, respectively. Experiments verified the effectiveness of the proposed algorithm.

2. SIGNAL MODEL

Consider an array containing two microphones to capture signals from one target speaker and K interfering speakers. The microphone

This work is supported by the National Key Research and Development Plan of China (No.2020AAA0140003), the National Natural Science Foundation of China (NSFC) (No.U21B2010, No.61831022, No.62101553, No.61971419, No.62006223, No.62276259, No.62201572, No.62206278), Beijing Municipal Science&Technology Commission,Administrative Commission of Zhongguancun Science Park(No.Z211100004821013), Open Research Projects of Zhejiang Lab (NO. 2021KH0AB06).

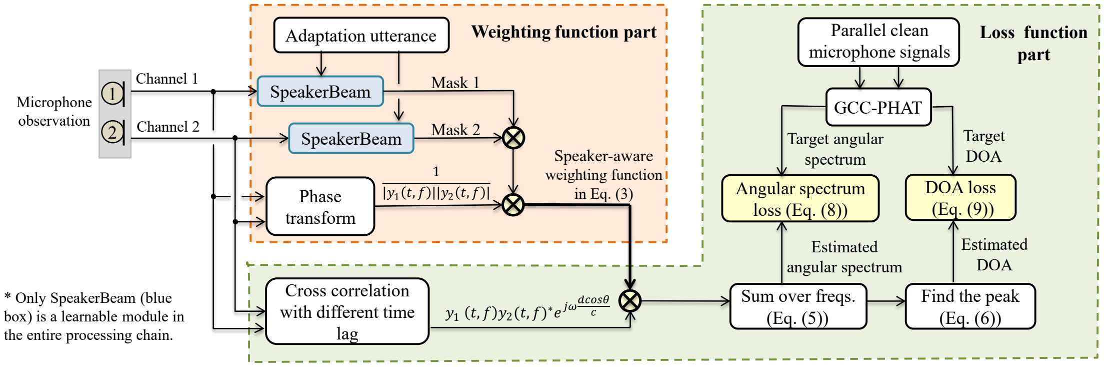
Microphone observation
1
2
Channel 1
Channel 2
Adaptation utterance
Weighting function part
SpeakerBeam
SpeakerBeam
Mask 1
Mask 2
Phase transform
$\frac{1}{|y_1(t,f)||y_2(t,f)|}$
Speaker-aware weighting function in Eq. (3)
Cross correlation with different time lag
$y_1(t,f)y_2(t,f)^*e^{j\omega\frac{d\cos\theta}{c}}$
Parallel clean microphone signals
Loss function part
GCC-PHAT
Target angular spectrum
Target DOA
Angular spectrum loss (Eq. (8))
DOA loss (Eq. (9))
Estimated angular spectrum
Estimated DOA
Sum over freqs. (Eq. (5))
Find the peak (Eq. (6))
* Only SpeakerBeam (blue box) is a learnable module in the entire processing chain.

Fig. 1. Overall scheme of GCC-speaker. **In the training phase**, the weighting function part (red dotted box) first outputs the speaker-dependent weighting, which is then fed into the loss function part (green dotted box). Gradients are backpropagated from the loss all the way back to the mask estimator (SpeakerBeam in the blue box). **In the inference phase**, we only need the speaker-dependent weighting function calculated from the weighting function part (red dotted box) and integrate it into the conventional GCC algorithm to find the target speaker's DOA.

observation vector $\mathbf{y}(t,f) = [y_1(t,f), y_2(t,f)]^T$ in the short-time Fourier transform (STFT) domain is given by

$$\mathbf{y}(t,f) = \mathbf{x}(t,f) + \sum_{k=1}^{K} \mathbf{i}_k(t,f) + \mathbf{v}(t,f), \quad (1)$$

where $t = 1, ..., T$ and $f = 1, .., F$ denote the time and frequency indices respectively, and $\mathbf{x}(t,f)$, $\mathbf{i}_k(t,f)$ and $\mathbf{v}(t,f)$ are the signal vectors of the target speaker, the k-th interfering speaker and the ambient noise respectively.

The GCC algorithm [2] can be applied to estimate the DOA at time index t, i.e.,

$$\hat{\theta}(t) = \arg\max_{\theta} \sum_{f=1}^{F/2} \Phi(t,f) y_1(t,f) y_2^*(t,f) e^{j\omega\frac{d\cos\theta}{c}}, \quad (2)$$

where superscript * denotes the complex conjugate, $j = \sqrt{-1}$ is the imaginary unit, c is the speed of sound, ω is the angular frequency, d is the microphone distance and $\Phi(t,f)$ is a weighting function. In order to discard the autocorrelation effect of speech, the GCC-PHAT algorithm [2] sets $\Phi(t,f) = \frac{1}{|y_1(t,f)||y_2^*(t,f)|}$ to whiten the source amplitude spectrum. To further improve the robustness of GCC to ambient noise $\mathbf{v}(t,f)$, a mask-based weighting term is applied to obtain $\Phi(t,f)$ [9, 18].

However, in multi-speaker scenarios, interfering speakers are non-negligible factors, while the existing GCC algorithms lack the ability to distinguish between the target speaker $\mathbf{x}(t,f)$ and the interfering speaker $\mathbf{i}_k(t,f)$ in Eq. (1) . Therefore, in the next section, we will design a speaker-dependent weighting function for the GCC algorithm to localize only the target speaker in multi-speaker scenarios.

3. THE PROPOSED GCC-SPEAKER

The scheme of the proposed GCC algorithm, named GCC-speaker, is shown in Fig. 1. It contains two parts: the weighting function part and the loss function part. The weighting function part is used for both the training and inference phases while the loss function part is only used for the training phase. We will describe these two parts in details below.

3.1. Weighting function part

Due to the approximate W-disjoint orthogonality [19] of speech signals in the time-frequency (TF) domain, in multi-speaker scenarios, there are still many TF units dominated by the target speaker, which is useful for the GCC algorithm to localize the target speaker. Therefore, we propose a speaker-dependent weighting function $\Phi_{\text{spk}}(t,f)$ for the GCC algorithm:

$$\Phi_{\text{spk}}(t,f) = \frac{M_1(t,f)M_2(t,f)}{|y_1(t,f)||y_2^*(t,f)|}, \quad (3)$$

where $M_i(t,f)$ is the speaker-dependent mask estimated from the i-th microphone. The denominator in Eq. (3) is used to equally emphasize all frequencies similar to the GCC-PHAT algorithm and the numerator in Eq. (3) is design to select the TF bins dominated by the target speaker for DOA estimation. A data-driven scheme for estimating $M_i(t,f)$ is utilized, which is based on the well-known SpeakerBeam for target speech extraction [17, 20, 21, 22],

$$M_i(t,f) = \text{SpeakerBeam}\left(|y_i(t,f)|, |a(t,f)|\right), \ i = 1, 2, \quad (4)$$

where $|y_i(t,f)|$ and $|a(t,f)|$ are the amplitude spectra of the i-th microphone signal and the adaptation utterance. The adaptation utterance is a speech segment containing only the target speaker and is crucial to inform the SpeakerBeam about the target speaker. We refer the reader to [17, 20] for the implementing details of SpeakerBeam.

A straightforward way to get the speaker-dependent weighting function $\Phi_{\text{spk}}(t,f)$ in Eq. (3) is to use the original SpeakerBeam to obtain the speaker-dependent mask and then bring it into Eq. (3) (this corresponds to the m-GCC-PHAT algorithm in the experiments). However, since SpeakerBeam is optimized to estimate the spectrum of the target speaker, it does not necessarily guarantee the optimal localization performance. In addition, it is known that low frequencies, which are occupied by the speech, actually yield low-resolution spatial spectrum for localization. Therefore, simply using the amplitude based weighting function can even degrade the performance especially in reverberant conditions. In order to solve this problem, we propose to tightly combine SpeakerBeam with the GCC algorithm and design a novel loss function to maximize the localization performance to optimize SpeakerBeam so as to guarantee the optimal $\Phi_{\text{spk}}(t,f)$.

3.2. Loss function part

In this section, rather than using the speech enhancement objective to optimize SpeakerBeam as in [17], we first obtain the weighting function through Eq. (3) and (4), and then integrate it into the GCC algorithm to obtain an estimated angular spectrum $\hat{b}(t,\theta)$,

$$\hat{b}(t,\theta) = \sum_{f=1}^{F/2} \Phi_{\text{spk}}(t,f)\Re\left\{y_1(t,f)y_2^*(t,f)e^{j\omega\frac{d\cos\theta}{c}}\right\}, \qquad (5)$$

where $\theta = [0, \Delta\theta, 2\Delta\theta, ..., \pi]$ with spatial resolution $\Delta\theta$ and $\Re\{.\}$ denotes real part [1].

In theory, the estimated DOA $\hat{\theta}(t)$ can be obtained by applying the argmax_θ operation on $\hat{b}(t,\theta)$. As the argmax_θ is not differential, which hinders the backpropagation. In order to successfully backpropagate gradients, we use the softargmax_θ operation [23] to obtain $\hat{\theta}(t)$ from $\hat{b}(t,\theta)$,

$$\hat{\theta}(t) = \underset{\theta}{\text{softargmax}}(\hat{b}(t,\theta)) = \sum_{\theta=0}^{\pi} \frac{\exp(\beta\hat{b}(t,\theta))}{\sum\limits_{\theta'=0}^{\pi}\exp(\beta\hat{b}(t,\theta'))}\theta, \qquad (6)$$

where β is a hyper-parameter controlling the smoothness of the softargmax_θ operation.

With $\hat{b}(t,\theta)$ and $\hat{\theta}(t)$ obtained from Eq.(5) and (6), we define a localization-related loss using MTL as

$$\mathcal{L} = \alpha\mathcal{L}_{\text{AS}} + (1-\alpha)\mathcal{L}_{\text{DOA}}, \qquad (7)$$

$$\mathcal{L}_{\text{AS}} = \sum_{\theta,t}\left(b_{\text{tgt}}(t,\theta) - \hat{b}(t,\theta)\right)^2, \qquad (8)$$

$$\mathcal{L}_{\text{DOA}} = \sum_{t}\left(\theta_{\text{tgt}}(t) - \hat{\theta}(t)\right)^2, \qquad (9)$$

where $b_{\text{tgt}}(t,\theta)$ and $\theta_{\text{tgt}}(t)$ are the angular spectrum and DOA estimated by applying GCC-PHAT to the parallel clean microphone signals (without interfering speakers), and they can be regarded as ground truth values. A hyper-parameter α is used in Eq. (7) to balance $\mathcal{L}_{\text{AS}}$ and $\mathcal{L}_{\text{DOA}}$. we experimentally found that α is of importance for network convergence. The choosing of α will be discussed in the section 4.3.

During the training phase, the gradients calculated from Eq. (7) can be backpropagated all the way back to the SpeakerBeam, so that the mask estimated by SpeakerBeam can yield a $\Phi_{\text{spk}}(t,f)$ guaranteeing the optimal localization. During the inference phase, we only need to apply the trained SpeakerBeam on each microphone to get $\Phi_{\text{spk}}(t,f)$ using Eq. (3) and (4), and then bring it into the conventional GCC algorithm by setting $\Phi(t,f) = \Phi_{\text{spk}}(t,f)$ in Eq. (2) to estimate the target speaker's DOA.

4. EXPERIMENTAL RESULTS

4.1. Data

We used the room impulse responses (RIRs) to convolve the utterances from the Wall Street Journal (WSJ) corpus [24] to generate 2-channel mixtures. The image method [25] was used to generate RIRs. There were 7138 utterances from 83 speakers for training, 410 utterances from 10 speakers for validation and 330 utterances from 10 speakers for test. For each mixture in the training set, we added one interfering speaker with SIR of 0 dB on average. The target speaker and the interfering speakers were randomly located in angles from 0° to 180°. We randomly picked an adaptation utterance from the utterances of the target speaker (different from the utterance in the mixture). The average length of the adaptation utterance was 10 s. We did not add ambient noise in this experiment. It will be our future work to perform target speaker localization in the scenario where the ambient noise and multiple interfering speakers coexist.

We create 2 training sets: one (denoted as TR1) with T60 = 200 ms and inter-microphone spacing of 20 cm, and the other (denoted as TR2) with T60 varying in [200 ms, 300 ms, 500 ms] and inter-microphone spacing varying in [10 cm, 15 cm , 20 cm]. To facilitate testing, we generated 20 test sets. The first 11 test sets with inter-microphone spacing of 20 cm varied only in SIR between -5 dB and 5 dB. The last 9 test sets with an average SIR of 0 dB varied in inter-microphone spacing between 10 cm and 25 cm and T60 between 200 ms and 500 ms. Moreover, the speakers in the test set did not appear in the training and validation sets.

4.2. Settings

The STFT frame size was 32 ms with 50% overlap, and the spatial resolution $\Delta\theta = 5^\circ$. We set $\beta = 0.1$ in Eq. (6). we followed [20] for setting SpeakerBeam. The main network of SpeakerBeam consisted of one LSTM layer, two fully connected layers with ReLU activation and one fully connected layer with a sigmoid activation (numbers of neurons: 257-1080-1024-257). The second layer of the main network was chosen as the adaptation layer and factorized into 30 sub-layers. The auxiliary network of SpeakerBeam consisted of two layers with a ReLU activation and one layer with a linear activation (numbers of neurons: 50-50-30). Before training GCC-speaker, we first pre-trained SpeakerBeam to minimize the cross-entropy loss w.r.t the ideal binary masks (IBM), because we found experimentally that the pre-training helps the network converge. The Adam optimizer [26] was used. The learning rate was set to 1e-3 initially and reduced by half when the validation loss stopped decreasing after two epochs. We used the mean absolute error (MAE) and frame-level accuracy with error tolerance of 5° as the metrics to evaluate the localization performance.

We compared the proposed GCC-speaker with (1)GCC-PHAT, (2)the speaker-dependent mask-based GCC-PHAT algorithm (denoted as m-GCC-PHAT) and (3)the target speaker localization algorithm proposed in [12] (denoted as m-cWMM). When using GCC-PHAT, we regarded the localization result of each frame as the target speaker's DOA due to the fact that GCC-PHAT is unable to distinguish the target speaker from the interfering speakers. Both m-GCC-PHAT and m-cWMM used SpeakerBeam to obtain the speaker-dependent mask, which is directly used by GCC-PHAT or the complex Watson mixture model (cWMM) [27] for localization. Different from GCC-speaker, the SpeakerBeam in these two algorithms was trained with the objective of recovering clean signals.

4.3. Results

We first evaluated the sensitiveness of the hyper-parameter α in Eq. (7) using the TR1 set and the test set with SIR=0 dB. We explore the different α from $[0, 0.01, 0.03, 0.05, 0.1, 0.3, 0.5, 1]$. As can be seen from Fig. 2, neither $\mathcal{L}_{\text{AS}}$ alone ($\alpha = 1$) nor $\mathcal{L}_{\text{DOA}}$ alone ($\alpha = 0$)

[1] Because $b(t,\theta)$ and $\Phi_{\text{spk}}(t,f)$ are theoretically real-valued, the $\Re\{.\}$ operation do not affect the result of Eq. (5). Moreover, $\Re\{.\}$ allows us to calculate the real-valued gradients for backpropagation.

Table 1. Results (MAE[°] / frame-level accuracy[%]) on the test sets for different array geometries and T60s.

Method	d=20cm			d=15cm			d=10cm		
	200ms	300ms	500ms	200ms	300ms	500ms	200ms	300ms	500ms
GCC-PHAT	14.7/71.3	11.1/77.6	10.3/78.8	15.1/70.3	11.6/76.6	10.3/78.4	16.2/66.7	12.8/72.6	11.9/74.2
m-GCC-PHAT	12.4/73.8	12.1/74.5	14.7/69.2	13.6/70.3	12.8/72.1	14.9/67.4	15.0/64.0	13.8/67.4	16.0/63.4
m-cWMM	12.5/73.8	12.2/74.4	14.9/68.9	13.6/70.8	13.0/71.9	15.7/66.5	14.9/66.3	14.2/67.0	16.9/62.2
GCC-speaker-TR1	**9.3/80.6**	**7.9/83.0**	10.3/78.3	11.2/76.3	9.5/79.9	11.1/77.6	13.0/72.2	11.3/76.0	13.1/73.1
GCC-speaker-TR2	10.4/78.2	8.7/80.7	**9.7/79.1**	**11.1/76.4**	**9.2/79.9**	**9.5/79.1**	**12.3/72.9**	**10.1/77.2**	**10.9/75.7**

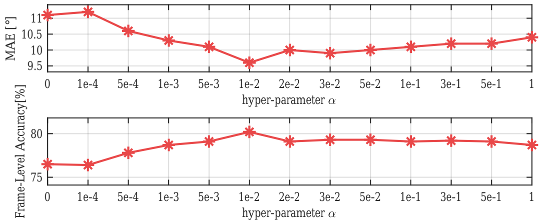

Fig. 2. The effect of the hyper-parameter α in Eq. (7).

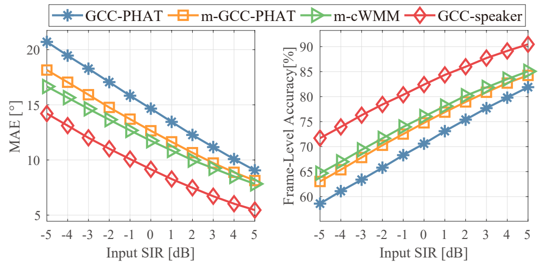

Fig. 3. MAE (left) and frame-level accuracy (right) as a function of the input SIR for the test sets.

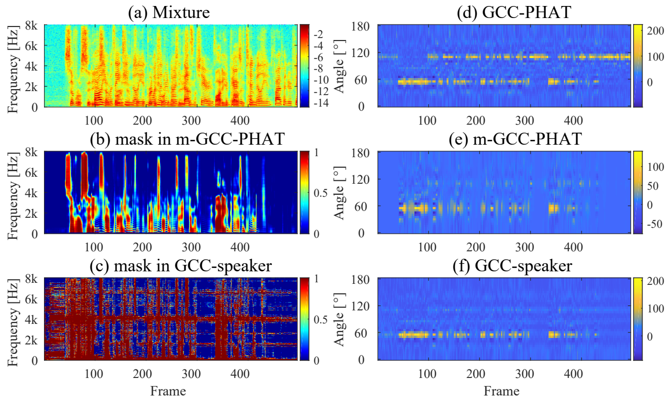

Fig. 4. An example of target speaker localization. The target speaker is at 55° and the interfering speaker is at 110°. (a) Amplitude spectrum. (b)-(c) Different masks. (e)-(f) Angular spectrums obtained by different algorithms.

in Eq. (7) guarantees the best performance, indicating that both $\mathcal{L}_{\text{AS}}$ and $\mathcal{L}_{\text{DOA}}$ are important for GCC-speaker. When $\alpha = 0.01$, relatively good results can be achieved. Therefore, in the following experiments, we kept $\alpha = 0.01$.

Next, we systematically evaluated the performances of the algorithms for different input SIRs using the TR1 set (see Fig. 3). As the input SIR increases, compared to GCC-PHAT, the superiority of m-GCC-PHAT and m-cWMM disappears. This is because using the mask with the objective of recovering clean signals may introduce distortion to the angular spectrum when suppressing interfering speakers (this will be further showed in Fig. 4). However, thanks to the end-to-end optimization, GCC-speaker significantly outperforms the comparative algorithms in different input SIRs.

In the following experiments, we systematically evaluated GCC-speaker for different array geometries and T60s using TR1 set and TR2 set. In this experiment, GCC-speaker-TR1 and GCC-speaker-TR2 represent the GCC-speaker models trained on the TR1 and TR2 set, respectively. As can be seen from Table 1, in the high-reverberation environment (T60=500ms), except for the GCC-speaker-TR2, all algorithms perform worse than GCC-PHAT. This may be due to the inaccurate mask estimation in unseen high-reverberation environments, which in turn affects the localization results of GCC. Moreover, Table 1 shows that GCC-speaker-TR1 can still perform better than m-GCC-PHAT and m-cWMM in the unseen environments, which indicates that GCC-speaker has a certain robustness for the array geometries and the T60s. However, GCC-speaker-TR2 can further improve the performance of GCC-speaker in the scenes where d=10 cm or d=15 cm.

Fig. 4 shows an example of target speaker localization. We can see from Fig. 4(e) and Fig. 4(f) that both m-GCC-PHAT and GCC-speaker can suppress the interfering speaker on the angular spectrum. However, m-GCC-PHAT may cause distortion of the target speaker's peak on the angular spectrum while GCC-speaker can well preserve the target speaker's peak. Moreover, as can be seen from Fig. 4(b) and Fig. 4(c), the mask in GCC-speaker does not restore the target signal on all TF units like the mask in m-GCC-PHAT. This may be due to the fact that not all the TF units are equally important in the target speaker localization task, which indicates that using the localization loss (e.q. (7)) instead of the speech enhancement loss is essential for a better localization performance.

5. CONCLUSIONS

In this paper, we propose a speaker-dependent weighting function for the GCC algorithm, so that GCC can localize only the target speaker in a multi-speaker scenario. Besides, the mask in the proposed weighting function is optimized with the objective on localization performance. Experiments show that the proposed algorithm significantly outperforms the existing target speaker localization algorithms. In the future we intend to improve the performance of target speaker localization in more complex acoustic environments.

6. REFERENCES

[1] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov, “A consolidated perspective on multimicrophone speech enhancement and source separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[2] Charles Knapp and Glifford Carter, “The generalized correlation method for estimation of time delay,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[3] Ralph Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[4] Yiteng Arden Huang, Jacob Benesty, and Jingdong Chen, “Time delay estimation and source localization,” in *Springer Handbook of Speech Processing*, pp. 1043–1063. Springer, 2008.

[5] DeLiang Wang and Jitong Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[6] Hong-Goo Kang, Michael Graczyk, and Jan Skoglund, “On pre-filtering strategies for the gcc-phat algorithm,” in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2016, pp. 1–5.

[7] Pasi Pertilä and Emre Cakir, “Robust direction estimation with convolutional neural networks based steered response power,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 6125–6129.

[8] Pasi Pertilä and Mikko Parviainen, “Time difference of arrival estimation of speech signals using deep neural networks with integrated time-frequency masking,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 436–440.

[9] Zhong-Qiu Wang, Xueliang Zhang, and DeLiang Wang, “Robust tdoa estimation based on time-frequency masking and deep neural networks.,” in *Interspeech*, 2018, pp. 322–326.

[10] Wei Xue, Ying Tong, Guohong Ding, Chao Zhang, Tao Ma, Xiaodong He, and Bowen Zhou, “Direct-Path Signal Cross-Correlation Estimation for Sound Source Localization in Reverberation,” in *Proc. Interspeech 2019*, 2019, pp. 2693–2697.

[11] Sunit Sivasankaran, Emmanuel Vincent, and Dominique Fohr, “Keyword-based speaker localization: Localizing a target speaker in a multi-speaker environment,” in *Interspeech 2018-19th Annual Conference of the International Speech Communication Association*, 2018.

[12] Ziteng Wang, Junfeng Li, and Yonghong Yan, “Target speaker localization based on the complex watson mixture model and time-frequency selection neural network,” *Applied Sciences*, vol. 8, no. 11, pp. 2326, 2018.

[13] Ulrik Kowalk, Simon Doclo, and Joerg Bitzer, “Signal-informed dnn-based doa estimation combining an external microphone and gcc-phat features,” *arXiv preprint arXiv:2206.05606*, 2022.

[14] Ofer Schwartz, Yuval Dorfan, Emanuël AP Habets, and Sharon Gannot, “Multi-speaker doa estimation in reverberation conditions using expectation-maximization,” in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2016, pp. 1–5.

[15] Shoko Araki, Hiroshi Sawada, Ryo Mukai, and Shoji Makino, “Doa estimation for multiple sparse sources with arbitrarily arranged multiple sensors,” *Journal of Signal Processing Systems*, vol. 63, no. 3, pp. 265–275, 2011.

[16] Koby Weisberg, Sharon Gannot, and Ofer Schwartz, “An online multiple-speaker doa tracking using the cappé-moulines recursive expectation-maximization algorithm,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 656–660.

[17] Kateřina Žmolíková, Marc Delcroix, Keisuke Kinoshita, Tsubasa Ochiai, Tomohiro Nakatani, Lukáš Burget, and Jan Černockỳ, “Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[18] Zhong-Qiu Wang, Xueliang Zhang, and DeLiang Wang, “Robust speaker localization guided by deep learning-based time-frequency masking,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 178–188, 2018.

[19] Ozgur Yilmaz and Scott Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Transactions on signal processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[20] Katerina Zmolikova, Marc Delcroix, Keisuke Kinoshita, Takuya Higuchi, Atsunori Ogawa, and Tomohiro Nakatani, “Speaker-aware neural network based beamformer for speaker extraction in speech mixtures.,” in *Interspeech*, 2017, pp. 2655–2659.

[21] Juan M Martín-Doñas, Jens Heitkaemper, Reinhold Haeb-Umbach, Angel M Gomez, and Antonio M Peinado, “Multi-channel block-online source extraction based on utterance adaptation,” *INTERSPEECH 2019, Graz, Austria*, 2019.

[22] Guanjun Li, Shan Liang, Shuai Nie, Wenju Liu, Meng Yu, Lianwu Chen, Shouye Peng, and Changliang Li, “Direction-aware speaker beam for multi-channel speaker extraction,” *Proc. Interspeech 2019*, pp. 2713–2717, 2019.

[23] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua, “Lift: Learned invariant feature transform,” in *European Conference on Computer Vision*. Springer, 2016, pp. 467–483.

[24] Douglas B Paul and Janet M Baker, “The design for the wall street journal-based csr corpus,” in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[25] Emanuel AP Habets, “Room impulse response generator,” *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, pp. 1, 2006.

[26] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[27] Dang Hai Tran Vu and Reinhold Haeb-Umbach, “Blind speech separation employing directional statistics in an expectation maximization framework,” in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 241–244.