Audio Visual Multi-Speaker Tracking with Improved GCF and PMBM Filter

Jinzheng Zhao[1], *Peipei Wu*[1], *Xubo Liu*[1], *Shidrokh Goudarzi*[1], *Haohe Liu*[1], *Yong Xu*[2], *Wenwu Wang*[1]

[1]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
[2]Tencent AI Lab, Bellevue, WA, USA

{j.zhao, W.Wang}@surrey.ac.uk, lucayongxu@tencent.com

Abstract

Audio and visual signals can be used jointly to provide complementary information for multi-speaker tracking. Face detectors and color histogram can provide visual measurements while Direction of Arrival (DOA) lines and global coherence field (GCF) maps can provide audio measurements. GCF, as a traditional sound source localization method, has been widely used to provide audio measurements in audio-visual speaker tracking by estimating the positions of speakers. However, GCF cannot directly deal with the scenarios of multiple speakers due to the emergence of spurious peaks on the GCF map, making it difficult to find the non-dominant speakers. To overcome this limitation, we propose a phase-aware VoiceFilter and a separation-before-localization method, which enables the audio mixture to be separated into individual speech sources while retaining their phases. This allows us to calculate the GCF map for multiple speakers, thereby their positions accurately and concurrently. Based on this method, we design an adaptive audio measurement likelihood for audio-visual multiple speaker tracking using Poisson multi-Bernoulli mixture (PMBM) filter. The experiments demonstrate that our proposed tracker achieves state-of-the-art results on the AV16.3 dataset.

Index Terms: speech separation, sound source localization, multiple-speaker tracking, audio-visual fusion

1. Introduction

Multi-speaker tracking aims at estimating the positions of multiple speakers based on sensor measurements. It plays an important role in a number of applications such as human-robot interaction [1], speech enhancement [2], and speaker diarization [3]. Audio and visual sensors have been used to improve the performance of speaker tracking systems by exploiting the complementarity between these two modalities. For instance, if the speakers are occluded by others or the illumination conditions are not good, audio signals can be used instead; if the audio information is affected by acoustic noise or the speakers are silent, we can turn to the visual data.

Audio measurements used in an audio-visual speaker tracking system can be obtained using a sound source localization (SSL) algorithm. One of the widely used SSL methods, global coherence field (GCF), which is based on Time Difference of Arrival (TDOA) estimation, has been employed to provide reliable measurements for speaker tracking [4] [5] [6]. The GCF feature is the summation of Generalized Cross Correlation with Phase Transform (GCC-PHAT) among all paired microphones. The peak on the GCF map indicates the position of the sound source. However, the estimated location may not be reliable if there are multiple speakers speaking concurrently due to the emergence of spurious peaks in the GCF map [7]. To deal with this problem, GCC-PHAT de-emphasis [7] has been proposed to adapt GCF in multi-speaker scenarios by calculating the GCF map again after localizing the dominant speaker. The updated GCF map is obtained by summing up a modified GCC-PHAT, which masks the time lags corresponding to the first speaker. Although this method can be used in the scenarios of multiple speakers, it has limited performance in localizing the non-dominant speakers, especially when the number of speakers increases [6].

In this paper, we propose to overcome the limitations of GCF by leveraging the techniques of speech separation, i.e., we separate the multi-speaker audio mixture into several single-speaker audio, and then calculate the GCF feature of individual sources for position estimations. We propose to use a phase-aware voicefilter for speech separation, built upon VoiceFilter [8], which is a method for target speech separation, giving promising performance in terms of signal to distortion ratio (SDR). VoiceFilter performs separation by estimating the magnitude spectrogram of the target speaker and reusing the phase mixture to reconstruct the waveform of the target speech (i.e., the phase of each separated audio is identical). However, the phase information is crucial for the GCF calculation, and should be correctly estimated. Therefore, in the phase-aware VoiceFilter, we propose to incorporate the phase estimation on the individual speaker inspired by [9, 10]. Phase-aware VoiceFilter contains a speaker recognition network and a target speaker separation network, where the former can produce unique embedding for a speaker, and the latter separates the target speaker from a mixture given the speaker embedding. By separating the overlapping audio into individual speech sources with different phases, the problem of multi-speaker SSL can be converted to single speaker SSL, which allows GCF to be adapted for the multi-speaker scenario.

Using the phase-aware VoiceFilter, the audio measurements, i.e. GCF, can be obtained more accurately, which can then be combined with visual measurements (such as a face bounding box obtained by a face detector), to improve the performance of an audio-visual tracking system. To fuse the audio and visual measurements, Bayesian-based filters, such as Particle filter (PF) [11], can be used, which is a sequential Monte Carlo algorithm approximating the state distribution by a number of weighted particles obtained by sequential importance sampling. However, PF cannot deal with the scenarios where the number of speakers is changing and unknown in different time steps. To deal with this issue, SMC-PHD filter [12] was proposed to estimate the varying number of speakers and their positions using audio and visual signals. Poisson multi-Bernoulli mixture (PMBM) filter was proposed in [13] [14] based on the conjugacy property that the predicted and updated distribution follows the same distribution. PMBM employs a Poisson point process to describe the distribution of undetected objects and employs a multi-Bernoulli mixture to describe the distribution of detected objects. PMBM outperforms other Bernoulli-based filters in terms of speed and ac-

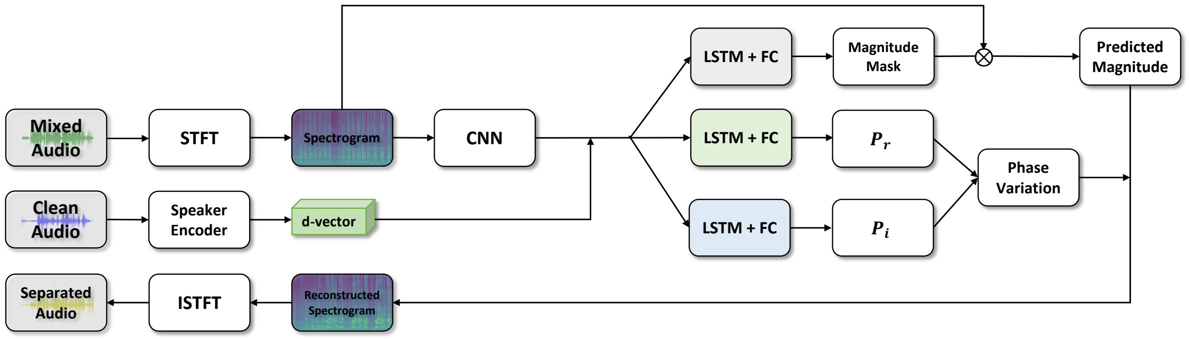


Figure 1: *The architecture of the phase-aware voicefilter*

curacy [15]. Due to its superiority, PMBM has been applied to tracking with visual information [16] or LiDAR signal [17]. We employ PMBM for audio-visual tracking, as in our previous work [18]. Different from [18], however, here we design different audio features and audio likelihood.

Our contributions in this paper are two-folds: (1) We propose a novel SSL method with the proposed phase-aware VoiceFilter, in order to improve the quality of audio measurements (i.e. GCF) for multi-speaker scenarios. (2) Based on this method, we design an adaptive audio measurements likelihood for audio-visual speaker tracking using the PMBM filter. Experimental results show that the proposed audio-visual tracker offers better performance than the baseline methods.

2. Proposed Method

We describe the generation of visual measurements, audio measurements, and the tracking framework in this section.

2.1. Generation of Visual Measurements

The face detector Dual-Shot Face Detector (DSFD) [19] is employed to generate visual measurements due to its promising performance in detecting faces accurately and robustly, which can output positions of bounding boxes and related confidence scores. More specifically, $\boldsymbol{b}_{k,i} = (x, y, w, h)^T$ denotes the i-th bounding box at time k, where (x, y) represents the bounding box's top left coordinates (subscripts i and k omitted for convenience) and (w, h) represents the width and height, respectively. We select the bounding boxes with confident scores above a predefined threshold and convert the coordinates of bounding boxes to the coordinates of the mouth position:

$$\boldsymbol{o}^v_{k,i} = \boldsymbol{L} \cdot \boldsymbol{b}_{k,i} \tag{1}$$

where $\boldsymbol{o}^v_{k,i}$ is employed as the i-th visual measurement (in superscript v) at time step k, $\boldsymbol{L} = [\boldsymbol{I}, \mathrm{diag}(0.5, 0.75)]$ is the matrix for transforming the coordinates of the bounding boxes to those of the mouth position [4].

2.2. Generation of Audio Measurements

We focus on the scenarios of two speakers. At first, phase-aware VoiceFilter is employed to extract the speech source of the targeted speaker conditioned on the d-vector, which is the unique embedding for each speaker. The d-vector is obtained by encoding a clean audio clip of the targeted speaker via a speaker encoder [20] pretrained on VoxCeleb2 [21] dataset. Then the GCF maps are calculated on the separated sources, respectively, to estimate the position of each speaker.

2.2.1. Phase-Aware VoiceFilter

VoiceFilter is a model for targeted speech separation, which consists of a speaker recognition network and a spectrogram masking network. However, VoiceFilter uses the phase mixture to reconstruct the separated waveform, thus the phase of each separated waveform is the same. Therefore, the positions of different speakers estimated by GCF are the same since localization by GCF depends on the phase difference between different microphones within a microphone array. To address this issue, we propose a phase-aware VoiceFilter, by incorporating phase prediction for the separated spectrogram.

We keep the speaker recognition network in VoiceFilter, which aims to produce a unique d-vector for each speaker. The original spectrogram masking network in VoiceFilter contains convolutional layers, LSTM layers and fully connected layers. It takes the magnitude spectrogram of mixed audio and d-vector as input, and generates a soft mask. This mask is multiplied with the magnitude spectrogram of the mixture audio to generate the speaker-oriented magnitude spectrogram. As shown in Figure 1, in addition to the soft mask prediction, we perform the phase variance prediction using two additional branches, from which the output P_{r} and P_{i} have the same shape as the magnitude mask. Then we calculate the phase variation θ as follows:

$$\cos\angle\theta = P_{\mathrm{r}} / \sqrt{P_{\mathrm{r}}^2 + P_{\mathrm{i}}^2} \tag{2}$$

$$\sin\angle\theta = P_{\mathrm{i}} / \sqrt{P_{\mathrm{r}}^2 + P_{\mathrm{i}}^2} \tag{3}$$

By combining the mixture phase with the estimated phase variation, the reconstructed magnitude spectrogram M_{r} and phase spectrogram M_{i} can be estimated as follows:

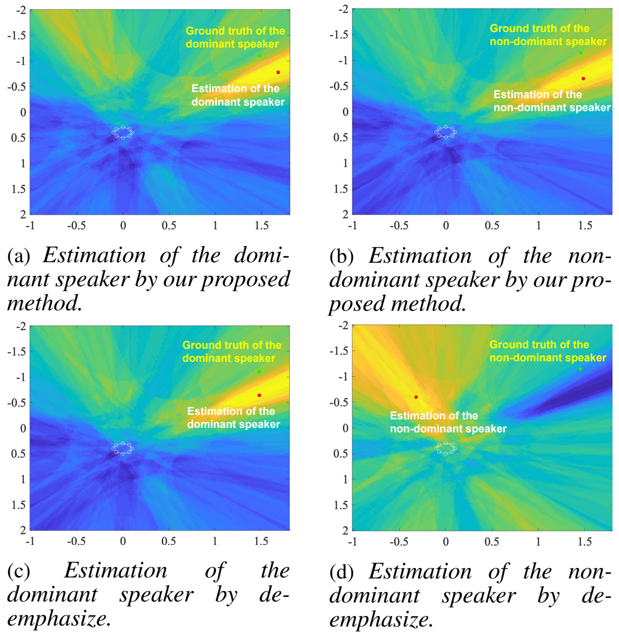
$$M_{\mathrm{r}} = M_{\mathrm{mag}} \cos(\angle M + \angle\theta) \tag{4}$$

$$M_{\mathrm{i}} = M_{\mathrm{mag}} \sin(\angle M + \angle\theta) \tag{5}$$

where $\angle M$ is the mixed phase and M_{mag} denotes the predicted magnitude, obtained by the multiplication of the predicted magnitude mask with the mixed magnitude. Finally, we obtain the separated audio by applying the inverse STFT on the complex spectrogram $M_{\mathrm{r}} + \mathrm{j} \cdot M_{\mathrm{i}}$, where j is the imaginary unit.

2.2.2. Global Coherence Field

Global coherence field (GCF) is a sound source localization method using audio signals from a microphone array. First, GCC-PHAT of audio signals from the j-th pair of microphones



(a) *Estimation of the dominant speaker by our proposed method.*

(b) *Estimation of the non-dominant speaker by our proposed method.*

(c) *Estimation of the dominant speaker by de-emphasize.*

(d) *Estimation of the non-dominant speaker by de-emphasize.*

Figure 2: *Estimation of positions of the dominant speaker and the non-dominant speaker in Seq02+Seq01 by our proposed method and the baseline method de-emphasize. In the GCF map, the intensity of the color denotes the possibility of the existence of the speaker. The white circle denotes the position of the microphone array. The estimations are calculated in the height of the speaker's mouths.*

in the array, i.e., $s_j \in S$ at time t, is calculated:

$$G_j(\tau, t)=\int_{-\infty}^{+\infty} \frac{T_{s_{j,1}}(t, f) T_{s_{j,2}}^*(t, f)}{\left|T_{s_{j,1}}(t, f)\right|\left|T_{s_{j,2}}^*(t, f)\right|} e^{j 2 \pi f \tau} d f \tag{6}$$

where τ is the inter-microphone time lag, f is frequency, $s_{j,1}$ and $s_{j,2}$ denote the two microphones of the j-th pair, T is the STFT and $*$ is the complex conjugate. GCF is the summation of GCC-PHAT over all microphone pairs. Following [4], we employ the speaker height z calculated by projecting the face bounding box to the 3D space to assist the calculation of GCF,

$$GCF(\boldsymbol{p}, t)=\frac{1}{|S|} \sum_{n=1}^{|S|} G_n\left(\tau_n(\boldsymbol{p} | z), t\right) \tag{7}$$

where $|S|$ is the number of microphone pairs, and $\boldsymbol{p}$ denotes possible positions over the entire space. The position $\boldsymbol{p}$ leading to the peak in the GCF map is regarded as the sound source position.

The phase-aware voicefilter is applied on the audio mixture from each microphone to obtain each individual speech source. The GCF maps are then calculated on each pair of separated sources, respectively. The positions $\boldsymbol{p}_1^a=(x_1, y_1, z_1)$ and $\boldsymbol{p}_2^a=(x_2, y_2, z_2)$ are estimated by picking peaks on GCF maps calculated on separated speech sources.

2.3. PMBM Filter

The PMBM filter can be used to estimate the number of speakers and the position of each speaker $\mathbf{x}=(x, y)$ at each time step, where (x, y) represents the location of the speaker. In each iteration, the speakers that are associated to measurements are defined as detected speakers, and the those not associated to measurements are defined as undetected speakers. Poisson point process $\mu(\cdot)$ is used to describe the distribution of undetected speakers $\mathbf{x}^u$ and multiple Bernoulli mixture $f(\cdot)$ is used to describe the detected speakers $\mathbf{x}^d$, where $\mathbf{x}^u$ and $\mathbf{x}^d$ are two disjoint subsets of existing speakers $\mathbf{x}$. The PMBM density $p_k(\cdot)$ which is used to represent the states of speakers, can be derived as the convolution of $\mu(\cdot)$ and $f(\cdot)$:

$$p_k(\mathbf{x})=\sum_{\mathbf{x}^u \uplus \mathbf{x}^d=\mathbf{x}} \mu_k\left(\mathbf{x}^u\right) f_k\left(\mathbf{x}^d\right) \tag{8}$$

2.3.1. *Prediction*

The predicted distribution $p_{k+1|k}\left(\mathbf{x}_{k+1}\right)$ can be calculated by the Chapman Kolmogorov equation:

$$p_{k+1|k}\left(\mathbf{x}_{k+1}\right)=\int \pi\left(\mathbf{x}_{k+1} \mid \mathbf{x}_k\right) p_{k|k}\left(\mathbf{x}_k\right) \delta \mathbf{x}_k \tag{9}$$

where $\pi\left(\mathbf{x}_{k+1} \mid \mathbf{x}_k\right)$ represents the prediction matrix. We suppose the speakers follow the uniform motion [22].

2.3.2. *Update*

The predicted distribution at time $k+1$ can be corrected with the measurement model $m\left(\mathbf{z}_{k+1} \mid \mathbf{x}_{k+1}\right)$:

$$p_{k+1|k+1}\left(\mathbf{x}_{k+1}\right)=\frac{m\left(\mathbf{z}_{k+1} \mid \mathbf{x}_{k+1}\right) p_{k+1|k}\left(\mathbf{x}_{k+1}\right)}{\int m\left(\mathbf{z}_{k+1} \mid \mathbf{x}_{k+1}^{\prime}\right) p_{k+1|k}\left(\mathbf{x}_{k+1}^{\prime}\right) \delta \mathbf{x}_k^{\prime}} \tag{10}$$

The measurement likelihood at time k is denoted by $m\left(\mathbf{z}_k \mid \mathbf{x}_k\right)$. For audio measurements $m\left(\mathbf{o}_k^a \mid \mathbf{x}_k\right)$, we design an adaptive audio likelihood based on our separation-before-localization method:

$$m\left(\mathbf{o}_k^a \mid \mathbf{x}_k\right) \propto \exp \left[-\left(\mathbf{o}_k^a-\mathbf{x}_k\right)^T \Sigma_a^{-1}\left(\mathbf{o}_k^a-\mathbf{x}_k\right)\right] \tag{11}$$

Let $\boldsymbol{p}_{1,k}^a$ and $\boldsymbol{p}_{2,k}^a$ denote the positions estimated from the separated audio and $\boldsymbol{p}_{m,k}^a$ is the position estimated from the mixed audio. They are converted to coordinates in the image plane $\boldsymbol{y}_{1,k}^a, \boldsymbol{y}_{2,k}^a, \boldsymbol{y}_{m,k}^a$ through the camera calibration information [23]. We also include the audio measurements $\boldsymbol{y}_{m,k}^a$ derived from the mixed audio in the audio likelihood in case the mixed audio is not well separated. The audio measurement $\mathbf{o}_k^a$ contains $\left\{\boldsymbol{y}_{1,k}^a, \boldsymbol{y}_{2,k}^a, \boldsymbol{y}_{m,k}^a\right\}$ if the corresponding GCF peak value $v_{i,k}^a(i=1,2,m)$ is beyond the threshold λ.

The visual likelihood follow Gaussian distribution centered at the estimated position $\mathbf{o}_k^v=\left\{\mathbf{o}_{k,1}^v, \mathbf{o}_{k,2}^v, \ldots, \mathbf{o}_{k,N}^v\right\}$ (N is the number of bounding boxes) calculated in Section 2.1:

$$m\left(\mathbf{o}_k^v \mid \mathbf{x}_k\right) \propto \exp \left[-\left(\mathbf{o}_k^v-\mathbf{x}_k\right)^T \Sigma_v^{-1}\left(\mathbf{o}_k^v-\mathbf{x}_k\right)\right] \tag{12}$$

The distribution of audio likelihood and visual likelihood is assumed to be independent. The audio and visual measurements are fused as follows:

$$m\left(\mathbf{z}_k \mid \mathbf{x}_k\right)=m\left(\mathbf{o}_k^a \mid \mathbf{x}_k\right) \cdot m\left(\mathbf{o}_k^v \mid \mathbf{x}_k\right) \tag{13}$$

3. Experiments

3.1. Dataset

We employed AV16.3 [25] dataset for performance evaluations. In Av16.3, there are two circular arrays with each containing 16 microphones to record audio, and three cameras to record the video. In the experiment, we focus on the scenarios of two-speaker tracking. The speech separation models tend to perform

Table 1: *Tracking Results. The results of AV-A-PF [22], MS-SMC-PHD [12] and AV-GLMB [24] are from [24].*

Sequence	Seq18			Seq19			Seq24			Seq25			Seq30			Avg
Camera	1^{st}	2^{nd}	3^{rd}	1^{st}	2^{nd}	3^{rd}	1^{st}	2^{nd}	3^{rd}	1^{st}	2^{nd}	3^{rd}	1^{st}	2^{nd}	3^{rd}	
AV-A-PF [22]	14.3	11.7	15.8	11.9	9.6	12.1	**10.0**	**8.9**	10.0	14.8	**7.7**	**8.9**	13.8	8.9	**10.3**	11.3
MS-SMC-PHD [12]	-	-	-	-	-	-	14.0	15.0	14.1	15.7	13.9	17.1	16.7	16.9	19.3	15.8
AV-GLMB [24]	15.7	10.9	**6.3**	15.3	11.6	**5.4**	16.5	10.6	**7.0**	17.7	10.8	10.7	14.8	10.4	15.7	12.0
Proposed	**9.6**	**10.2**	8.6	**10.1**	**9.2**	7.2	12.1	11.9	8.3	**9.8**	10.9	12.0	**11.3**	**7.8**	19.5	**10.6**

Table 2: *MSE results for opposite gender audio mixtures.*

Sequence	02 + 01	02 + 03	02 + 12	02 + 15	18	Avg
De-empha[7]	1.08	1.11	0.98	1.10	**0.69**	0.99
Proposed	**0.75**	**0.92**	**0.93**	**1.10**	0.71	**0.88**

Table 3: *MSE results for the same gender audio mixtures.*

Sequence	19	24	25	30	Avg
De-empha[7]	**0.76**	**0.67**	1.43	0.99	**0.90**
Proposed	0.83	1.42	**0.63**	**0.84**	0.93

better in opposite gender audio mixture [26]. To demonstrate the advantages of the proposed separation-before-localization method, we evaluate the model performance on opposite gender audio mixtures and the same-gender audio mixtures, respectively. For opposite gender audio mixtures, we select sequence 18. In addition, we create additional sequences by summing up the audio of sequence 02 and sequence 01 (02+01), sequence 02 and sequence 03 (02+03), sequence 02 and sequence 12 (02+12), sequence 02 and sequence 15 (02+15), where sequence 02 is one female speaking and other sequences are one male speaking. For the same-gender audio mixtures, we select sequences 19, 24, 25 and 30.

3.2. Implementation Details

For phase-aware VoiceFilter, the CNN block in Figure 1 has eight convolutional layers. Each layer is preceded by the ZeroPad2d layer and followed by BatchNorm2d and ReLU layer. The LSTM block has bidirectional architecture with 5064 dimensional input layer and 400 dimensional hidden layer. The model is trained on Librispeech [27] dataset with 200,000 steps. To obtain the d-vector, we extract audio clips from single speaker sequences which has the same speaker as in the multiple speakers sequences. The audio clips are input to the pretrained speaker encoder [20] to get the d-vector. For face detecting, the predicted bounding boxes by DSFD whose confidences are above 0.8 are reserved for visual measurements.

3.3. Analysis of the Quality of Audio Measurements

We use Mean Square Error (MSE) to measure the reliability of audio measurements. We implement the de-emphasis method [7] and compare it with our proposed method. The results on opposite gender audio mixtures are listed in Table 2. For each sequence, the average MSE in localizing the dominant speaker and the non-dominant speaker is reported. The average MSE (Avg) over all sequences is shown in the last column. It is shown that our proposed method outperforms the de-emphasis method. We also visualize the GCF map in Figure 2. In this case, the two speakers are standing very closely (One stands at $(1.452m, -1.145m)$ and the other at $(1.482m, -1.106m)$). When estimating the position of the dominant speaker (Figure 2a and Figure 2c), the proposed method shows comparative performance with the baseline method. When estimating the position of the second speaker (Figure 2d), the estimation by the de-emphasis method deviates substantially from the ground truth. The reason is that de-emphasis masks the time lags corresponding to the dominant speaker. However, as the two speakers are very close, the mask used can remove the positional cues corresponding to the non-dominant speaker. The spurious peak caused by reverberation or background noise on the left in Figure 2d emerges. Our method can perform robustly in this circumstance (Figure 2b). The MSE results on the same gender audio mixtures are shown in Table 3. The performance of the proposed method is comparative to the baseline, and is not as good as that on opposite gender audio mixtures. The reason is that the mixed audio is not well separated by VoiceFilter as the model tends to have more difficulty in separating the same gender mixture than opposite gender mixture [26].

3.4. Analysis of Tracking Results

We compare our proposed tracker on sequences of multiple speakers with state-of-the-art methods, AV-A-PF [22], MS-SMC-PHD [12] and AV-GLMB [24]. We test the proposed tracker ten times and the average results are calculated. The PMBM tracker is initialized with measurements at the starting frame. We also use MSE to evaluate the tracker's performance and the results are shown in Table 1. The last column (Avg) demonstrates the overall performance of each tracker. In AV-A-PF, the number of targets is known, and is initialized with ground truth, which reduces the task difficulties. Our proposed tracker does not need to know the number of targets and can deal with the scenarios of a time-varying number of targets. It gives the lowest tracking errors as compared to the baseline methods.

4. Conclusions

We have presented a novel sound source localization method with the assistance of the proposed phase-aware VoiceFilter, which overcomes the limitation of the traditional GCF method. This method provides better localization performance as compared to the baseline method on opposite gender audio mixtures and achieves performance that is comparable to the baseline method on the same gender audio mixtures. The proposed tracker using visual and audio measurements achieves state-of-the-art results on the AV16.3 dataset.

5. Acknowledgements

This research is supported by Tencent AI Lab Rhino-Bird Gift Fund and University of Surrey. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

6. References

[1] S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "Audio-visual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 882–894, 2010.

[2] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[3] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1086–1099, 2017.

[4] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "Multi-speaker tracking from an audio-visual sensing device," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2576–2588, 2019.

[5] Y. Li, H. Liu, and H. Tang, "Multi-modal perception attention network with self-supervised learning for audio-visual speaker tracking," *arXiv:2112.07423*, 2021.

[6] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "Audio-visual tracking of concurrent speakers," *IEEE Transactions on Multimedia*, 2021.

[7] A. Brutti, M. Omologo, and P. Svaizer, "Localization of multiple speakers based on a two step acoustic map analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 4349–4352.

[8] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," *arXiv:1810.04826*, 2018.

[9] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang, "Decoupling magnitude and phase estimation with deep resunet for music source separation," *arXiv:2109.05418*, 2021.

[10] H. Liu, Q. Kong, and J. Liu, "CWS-PResUNet: Music source separation with channel-wise subband phase-aware ResUNet," *arXiv:2112.04685*, 2021.

[11] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[12] V. Kılıç, M. Barnard, W. Wang, A. Hilton, and J. Kittler, "Mean-shift and sparse sampling-based SMC-PHD filtering for audio informed visual speaker tracking," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2417–2431, 2016.

[13] J. L. Williams, "Marginal multi-Bernoulli filters: RFS derivation of MHT, JIPDA, and association-based MeMBer," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 51, no. 3, pp. 1664–1687, 2015.

[14] Á. F. García-Fernández, J. L. Williams, K. Granström, and L. Svensson, "Poisson multi-bernoulli mixture filter: direct derivation and implementation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 4, pp. 1883–1901, 2018.

[15] Y. Xia, K. Granstrcom, L. Svensson, and Á. F. García-Fernández, "Performance evaluation of multi-bernoulli conjugate priors for multi-target filtering," in *IEEE 20th International Conference on Information Fusion (Fusion)*, 2017, pp. 1–8.

[16] S. Scheidegger, J. Benjaminsson, E. Rosenberg, A. Krishnan, and K. Granström, "Mono-camera 3d multi-object tracking using deep learning detections and pmbm filtering," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, 2018, pp. 433–440.

[17] S. Pang and H. Radha, "Multi-object tracking using poisson multi-bernoulli mixture filtering for autonomous vehicles," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7963–7967.

[18] J. Zhao, P. Wu, X. Liu, Y. Xu, L. Mihaylova, S. Godsill, and W. Wang, "Audio-visual tracking of multiple speakers via a PMBM filter," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 5068–5072.

[19] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang, "DSFD: dual shot face detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5060–5069.

[20] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4879–4883.

[21] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv:1806.05622*, 2018.

[22] V. Kılıç, M. Barnard, W. Wang, and J. Kittler, "Audio assisted robust visual tracking with adaptive particle filtering," *IEEE Transactions on Multimedia*, vol. 17, no. 2, pp. 186–200, 2014.

[23] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

[24] S. Lin and X. Qian, "Audio-visual multi-speaker tracking based on the glmb framework." in *INTERSPEECH*, 2020, pp. 3082–3086.

[25] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16. 3: An audio-visual corpus for speaker localization and tracking," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2004, pp. 182–195.

[26] K. Wang, F. Soong, and L. Xie, "A pitch-aware approach to single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 296–300.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015, pp. 5206–5210.