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/ICASSP49357.2023.10095549

INCORPORATING LIP FEATURES INTO AUDIO-VISUAL MULTI-SPEAKER DOA ESTIMATION BY GATED FUSION

Ya Jiang[1], *Hang Chen*[1], *Jun Du*[1,*], *Qing Wang*[1], *Chin-Hui Lee*[2]

[1]University of Science and Technology of China, Hefei, Anhui, P. R. China
[2]Georgia Institute of Technology, Atlanta, GA. USA
yajiang@mail.ustc.edu.cn, ✉jundu@ustc.edu.cn

ABSTRACT

The audio-visual direction of arrival (DOA) estimation has demonstrated superior performance recently. In this paper, we present a novel audio-visual multi-speaker DOA estimation network, which for the first time incorporates multi-speaker lip features to adapt the complex overlapping and noisy scenarios. Firstly, we encode the multi-channel audio features, the reference angles and the lip Regions of Interest (RoIs) detected from the video respectively to acquire high-level representations. Then the multi-modal embeddings of audio, speaker angles and lips are fused by a tri-modal gated fusion module to balance their contributions to the output. The fused embedding is sent to the backend network to obtain the accurate DOA estimation with the combination of the predicted speaker angular vectors and the speaker activities. Experimental results show that our proposed approach can reduce the localization error by 73.48% compared to the previous work on the 2021 Multi-modal Information based Speech Processing (MISP) Challenge corpus. Meanwhile, the high accuracy and stability of localization results demonstrate the robustness of the proposed model in multi-speaker scenarios.

***Index Terms*—** Multi-speaker DOA estimation, audio-visual sound source localization, lip embedding, multi-modal information, gated fusion

1. INTRODUCTION

The direction of arrival (DOA) estimation aims to locate spatial positions of one or more sources relative to the microphone array utilizing captured speech signals. The DOA estimation has been applied widely in speech enhancement [1], teleconferencing [2], automatic speech recognition [3], and other acoustic signal processing areas.

Traditional array signal processing techniques have developed a series of DOA estimation methods over decades, such as the time delay of arrival (TDOA) based methods utilizing generalized cross-correlation (GCC) [4]; subspace based approaches represented by multiple signal classification (MUSIC) [5]; the signal synchronization based methods like the steered response power with phase transform (SRP-PHAT) [6]; the blind identification of impulse response based methods such as the adaptive eigenvalue decomposition (AED) [7]. Traditional methods have made impressive progress, but they rely heavily on ideal assumptions such as the white noise and the high SNR, which are difficult to maintain in real scenarios.

Lately, many researchers have explored deep learning based approaches to improve the robustness of systems in adverse scenes. Typically, the inputs of neural networks can be the GCC-PHAT (Phase Transform) Patterns [8], the raw waveform [9], the magnitude and phase parts of frequency domain features [10], etc. And the outputs can be positions classified with a predefined angular resolution [11], or angles predicted by location regression [12]. Most recently, the advanced transformer-based model [13] was proposed to handle DOA estimation problem in multi-source scenes. Additionally, a DNN based steered response power (SRP) method [14] utilizing spatial-temporal context information was published for multiple moving sources. Although the accuracy of DOA estimation has been improved thanks to the powerful learning ability of neural networks, overlaps and heavy noises still pose great challenges.

Considering that humans naturally integrate auditory and visual stimuli to obtain a profound understanding of the real world, some studies have proposed to combining audio and visual information for better insight into localizing sound sources. A novel real-time audio-visual system to localize all active speakers with a full 360° was proposed in [15]. [16] adopted a novel deep generative beamformer (DGB) to incorporate a deep learning process with the conventional SRP-PHAT beamforming for audio-visual sound source localization. A multi-speaker DOA estimation method utilizing the reference angular vectors detected from the video as an assistance to audio was proposed in our previous work [17], which is based on the transformation between the pixel coordinate system and the camera coordinate system according to the pin-hole camera model. With the help of the reference angles from the video, this work addressed the label permutation problem in multi-speaker DOA estimation.

In this paper, we construct an advanced audio-visual framework based on our previous work [17] to deeply mine the rich information from the video to overcome the difficulties of high overlapping ratios and heavy noises in real scenes. Specifically, our main contributions are: Firstly, we propose to incorporate lip feature of speakers into the audio-visual DOA task for the first time, which provides helpful information for the active speaker prediction in DOA estimation. Then we propose a multi-modal gated fusion method to balance the contributions of multiple embeddings to the network output. Besides, a speaker-wise loss function is presented to jointly optimize the predicted speaker active probabilities and the corresponding angular vectors. A series of audio-visual ablation experiments are designed to analyze the assistance and supplement of the visual modality to the audio modality, and evaluated on the MISP2021 corpus. Experimental results demonstrate that through combining audio features, speaker angular vectors and lip features, our proposed multi-modal DOA estimation network shows significantly improved performances with a relative localization error reduction of 73.48% compared to the baseline [17] with high accuracy and robustness.

The remainder of this paper is organized as follows: Section 2 introduces the overall architecture of our proposed audio-visual DOA network. Then we describe the experimental setups, results, and analysis in Section 3. Finally, we draw a conclusion in Section 4.

[*] corresponding author

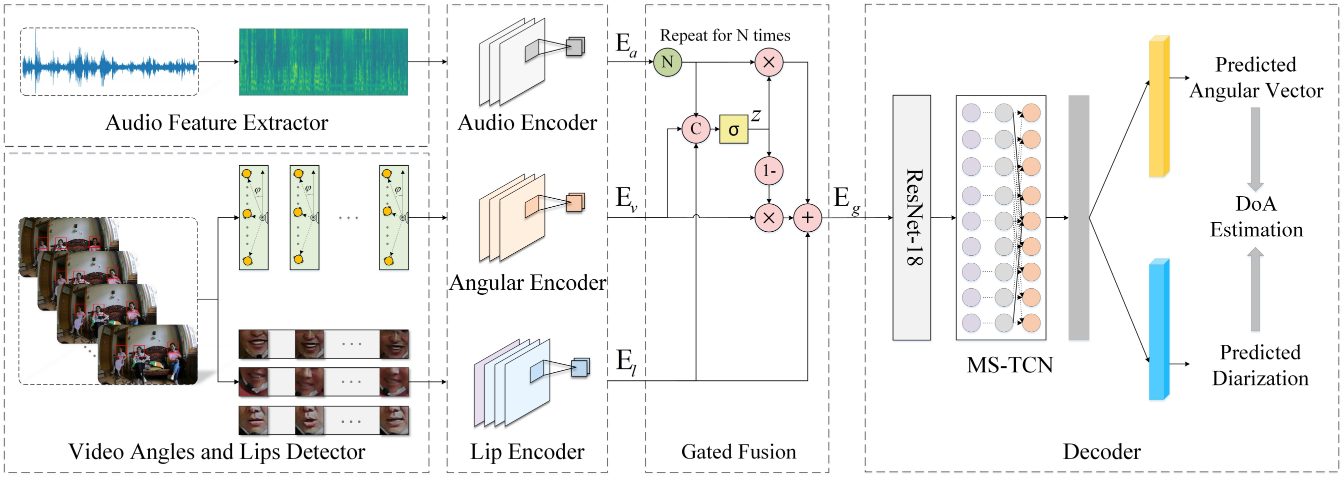

Fig. 1. The flowchart of the proposed audio-visual DOA framework.

2. PROPOSED METHOD

Fig. 1 illustrates the overall flowchart of the proposed multi-modal DOA estimation, which is a joint task of speaker localization and diarization. We first encode audio features and video speaker angles to acquire the audio embedding and the angular embedding respectively, and employ a pre-trained lip-reading model to extract the lip embedding simultaneously. Then the gated fusion module takes multi-modal embeddings as inputs and the decoder combines the predicted speaker angles and posterior probabilities to obtain the final DOA estimation.

2.1. Audio encoder

In the process of audio signals, we apply Mel filter bank (FBANK) features as the input features. Specifically, we perform a 64-dimensional FBANK extractor on the 6-channel audio with a window size of 20 ms and a window shift of the same length. After time-frequency analysis, the 6-channel FBANK features are fed into an 18-layer ResNet as our audio embedding extractor f_a. The process can be formulated as:

$$\mathbf{E}_a = f_a\left(\mathbf{F}_a\right) \tag{1}$$

where $\mathbf{E}_a = \left\{\boldsymbol{e}_a^1, \boldsymbol{e}_a^2, \cdots, \boldsymbol{e}_a^{T_a}\right\}, \boldsymbol{e}_a^t \in \mathbb{R}^{D_a}$ is the audio embedding and $\mathbf{F}_a \in \mathbb{R}^{C \times F_{\mathrm{mel}} \times T_a}$ is the FBANK features. And T_a is the number of audio frames, C is the number of audio channels, F_{mel} and D_a are the dimensions of the FBANK features and the audio embedding, respectively.

2.2. Angular encoder

To acquire the oracle angular information from the video modality, we utilize an off-the-shelf face and lip detector based on YoLo-v5 algorithm[1] to locate the face and lip Regions of Interest (RoIs) of each person in the video, where lip RoIs are at almost the center of face RoIs, regarded as the sound source locations. In this way, we accurately acquire the pixel coordinates of the sound sources in each video frame. Sequentially, the spatial annotation method [17] is implemented to transfer coordinates of the located sound sources in the pixel coordinate system to the camera coordinate system, and the corresponding azimuth angles relative to the camera are calculated as a result. We adopt the azimuth angles annotated from the video as the oracle DOA information in our proposed audio-visual DOA estimation model. Thus, we can compute the n-th speaker's azimuth angles in one sample as $\Phi_n = \{\phi_{n,1}, \phi_{n,2}, \cdots, \phi_{n,T_v}\}$, where T_v is the number of video frames. During encoding, we first express the azimuth angles as the angular vectors for better representation $\mathbf{X}_{v_n} = \{\boldsymbol{x}_{n,1}, \boldsymbol{x}_{n,2}, \cdots, \boldsymbol{x}_{n,T_v}\}$, where $\boldsymbol{x}_{n,t} = (\cos\phi_{n,t}, \sin\phi_{n,t})$. Then the T_v-frame angular vectors of N speakers (N is the number of the speakers detected in the video) are encoded to obtain the angular embedding by an 18-ResNet network f_v formulated as:

$$\mathbf{E}_v = f_v\left(\mathbf{X}_v\right) \tag{2}$$

where $\mathbf{E}_v = \left\{\boldsymbol{e}_v^1, \boldsymbol{e}_v^2, \cdots, \boldsymbol{e}_v^{T_v}\right\}, \boldsymbol{e}_v^t \in \mathbb{R}^{N \times D_v}$. And D_v is the dimension of the angular embedding for each person.

2.3. Lip encoder

The oracle angular information calibrated from the video is quite useful for DOA estimation, providing an accurate angular reference. Moreover, the performance of locating the active person only with audio signals is not satisfactory, because of the multiple overlapping sound sources and background noises in real scenes, such as the TV noise in MISP2021. In view of this, we crop the lip Regions of Interest (RoIs) of the target speakers as assistance while detecting their azimuth angles in Section 2.2. For the frames of missing lip detection, we search for the lip RoIs from the nearest non-empty frames before and after them, then perform linear interpolations between the non-empty frames. Finally we save the maximum range size of detected lip RoIs for each speaker within each sample duration and resize all the cropped lip RoIs to a fixed size of 96 × 96 × 3. After this, we take the grayed-out lip RoIs of N speakers detected in the video $\mathbf{L}_v = (\boldsymbol{l}_1, \boldsymbol{l}_2, \cdots, \boldsymbol{l}_{T_v})$ as the input to calculate the lip embedding $\mathbf{E}_l$ utilizing a lip-reading model pre-trained on clustered triphone [18] as our lip encoder f_l, which consists of a spatio-temporal convolution (including a convolution layer with 64 3D-kernels, a batch normalization) followed by an 18-layer ResNet. We can formulate the process as:

$$\begin{aligned}\mathbf{E}_l &= f_l\left(\mathbf{L}_v\right)\\ &= \mathrm{ResNet\,18}\left(\mathrm{BN}\left(\mathrm{Conv}_{3D}\left(\mathbf{L}_v\right)\right)\right)\end{aligned} \tag{3}$$

where $\mathbf{E}_l = \left\{\boldsymbol{e}_l^1, \boldsymbol{e}_l^2, \cdots, \boldsymbol{e}_l^{T_v}\right\}, \boldsymbol{e}_l^t \in \mathbb{R}^{N \times D_l}$. And D_l is the dimension of the lip embedding for each person.

[1] https://github.com/ultralytics/yolov5

2.4. Tri-modal gated fusion module

To solve the temporal alignment problem of audio and video embeddings, we repeat two times for the angular embedding and the lip embedding of the video. Besides we repeat N times for the audio embedding before fusing it with the others.

Although simple element-wise addition, concatenation and dot operations can fuse multi-embedding vectors, they are not highly efficient in capturing high-level correlations among them. Moreover, with the help of reference angular vectors, the audio and speaker's lips provide more contributions to active speaker prediction rather than localization. For these reasons, we incorporate the gated fusion module into the DOA model to acquire high-level associations among multiple embeddings. Specifically, we utilize a tri-modal gated fusion module modified by the bimodal approach proposed in the gated multimodal unit (GMU) network [19] to learn the fused embedding $\mathbf{E}_g$, which can be described as:

$$\begin{aligned} \boldsymbol{z}_t &= \sigma\left(\mathbf{W}_z \cdot \left[\boldsymbol{e}_a^t, \boldsymbol{e}_v^t, \boldsymbol{e}_l^t\right]\right) \\ \boldsymbol{e}_g^t &= \boldsymbol{z}_t * \boldsymbol{e}_a^t + (1-\boldsymbol{z}_t) * \boldsymbol{e}_v^t + \mathbf{W}_g \cdot \boldsymbol{e}_l^t \end{aligned} \tag{4}$$

where $\mathbf{E}_g = \left\{\boldsymbol{e}_g^1, \boldsymbol{e}_g^2, \cdots, \boldsymbol{e}_g^{T_a}\right\}, \boldsymbol{e}_g^t \in \mathbb{R}^{N \times D}$ is the fused embedding, and D is the dimension of the fused embedding for each person. And $\boldsymbol{e}_a^t \in \mathbf{E}_a$, $\boldsymbol{e}_v^t \in \mathbf{E}_v$ and $\boldsymbol{e}_l^t \in \mathbf{E}_l$ are the vectors of the audio embedding, the angular embedding and the lip embedding at time t, respectively. Meanwhile $\Theta = \{\mathbf{W}_z, \mathbf{W}_g\}$ are the parameters to be learned and $[\cdot,\cdot,\cdot]$ denotes the concatenation operator. And σ stands for the Sigmoid function.

During multi-modal gated fusing, the fusion module takes three kinds of information as the input, and each $\boldsymbol{e}_i$ corresponds to an embedding vector associated with modality i. With a gate neuron σ, the DOA network can control the contributions of the features calculated from $\boldsymbol{e}_i$ to the overall output $\boldsymbol{e}_g$ of the fusion module.

2.5. Decoder

The fused embedding is fed into an 18-layer ResNet, followed by a Multi-Scale Temporal Convolution Network (MS-TCN) [20] as the temporal sequential modeling module. The outputs of MS-TCN go through two branches respectively. One branch is a fully-connected (FC) layer followed by the L2 normalization, corresponding to the predicted angular vectors whose magnitudes are normalized as 1. And the other is a FC layer followed by a Sigmoid activation, corresponding to the predicted active posterior probabilities. During decoding, we binarize the active posterior probabilities by choosing a threshold to select the corresponding speaker angular vectors as the final DOA estimation.

2.6. Optimization

The total speaker-wise BceCosLoss of the network is the sum of binary cross entropy (BCE) loss of speaker activities and the speaker-wise Cosine loss of DOA estimation described as follows:

$$\mathcal{L} = \mathcal{L}_{\text{bce}}(\hat{\gamma}, \gamma) + \mathcal{L}_{\cos}(\hat{\boldsymbol{x}}, \boldsymbol{x}, \hat{\gamma}) \tag{5}$$

where $\mathcal{L}_{\text{bce}}(\hat{\gamma}, \gamma)$ is the BCE loss of the predicted active posterior probabilities $\hat{\gamma}$ of target speakers and the ground-truth activities γ, as well as $\mathcal{L}_{\cos}(\hat{\boldsymbol{x}}, \boldsymbol{x}, \hat{\gamma})$ is the speaker-wise Cosine loss of the predicted angular vectors $\hat{\boldsymbol{x}} = \left(\cos\hat{\phi}, \sin\hat{\phi}\right)$ and the corresponding ground-truth $\boldsymbol{x} = (\cos\phi, \sin\phi)$, with $\hat{\phi}$ and ϕ being the predicted angles and the corresponding ground-truth. We multiply the Cosine loss with the predicted speaker active posterior probabilities as a speaker-wise weighting function, instead of optimizing the BCE loss and the Cosine loss independently, calculated as:

$$\mathcal{L}_{\cos}(\hat{\boldsymbol{x}}, \boldsymbol{x}, \hat{\gamma}) = \hat{\gamma}\left(1 - \boldsymbol{cos}(\hat{\boldsymbol{x}}, \boldsymbol{x})\right) \tag{6}$$

where $\boldsymbol{cos}(\hat{\boldsymbol{x}}, \boldsymbol{x})$ is the cosine distance between $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}$. We only computed the BceCosLoss for frames with more than one labelled active speakers.

We minimize the loss function by adopting Adam optimizer [21] for 40 epochs with an initial learning rate of 3e-4 and a weight decay of 10e-4, and utilizing cosine scheduler [22].

3. EXPERIMENTS

3.1. Experimental setup

We conduct our experiments on the far-field recordings of the MISP2021 corpus [23]. The sample rate of the audio is 16 kHz, and the video is recorded at 25 frames per second. We focus on the azimuth angles because of the same heights of the sitting speakers and the linear microphone array, whose midpoint coincides with the origin of the camera. The multi-modal embeddings and the fused embedding have the same dimensions of 512. Meanwhile, the output DOA frame rate is the same as the video frame rate, i.e., 40 ms.

We report the mean absolute DOA Error for evaluation. Given a predicted angular vector list $\hat{\boldsymbol{x}}$ and the corresponding ground truth list $\boldsymbol{x}$, we apply the Hungarian algorithm [24] to solve the assignment problem. Then the absolute DOA Errors are computed as the cosine distances of the matched pairs $\boldsymbol{cos}(\hat{\boldsymbol{x}}, \boldsymbol{x})$. For mismatched pairs, we employ a punish angular offset of the maximum angular range of sources appearing at that frame. We find the best model with the lowest DOA Error on MISP2021 development set (DEV) and apply it to MISP2021 evaluation set (EVAL) to calculate the final DOA Error. We also report the accuracy of DOA estimation where the output is considered true positive only when the DOA Error is under a threshold of 20°, and the F1-Score calculated with the precision and recall of the correct predicted frames in each sample.

We design several variants of different combinations among audio features, speaker angles and lips for our proposed audio-visual DOA estimation model, i.e, the AV(AL) DOA model. The inputs are demonstrated as Table 1, specifically, A, AV(A), AV(L) and V(AL) denote the Audio-only model, the Audio-Visual (Angles) model, the Audio-Visual (Lips) model and the Video-only model respectively. The networks which receive angular vectors as a reference, including AV(A), V(AL) and AV(AL) models, output the angular differences between the predicted angles and ground truth, while the others directly regress the angles. We utilize the permutation invariant training (PIT) strategy [17] to align the speaker labels in A DOA model and adopt a bi-modal gated fusion without the parameter $\mathbf{W}_g$ for the third embedding in AV(A), AV(L) and V(AL) DOA models.

3.2. Results and analysis of ablation experiments

We presented the localization performances of DOA estimation models in Table 1. The A DOA model performed poorly with a high DOA Error in multi-speaker scenes. The AV(A) DOA model as well as the baseline model reduced the DOA Error with the assistance of the angular reference. However it is not very helpful to determine the activity of speakers as seen from the low F1-Score of 0.59 in the baseline model. On the other hand, the AV(L) DOA model effectively improved the F1-Score to 0.9 utilizing the speaker lips as an auxiliary to audio. Yet, without the video angular reference, the network cannot regress the speaker angles exactly, which leads to

Table 1. Specific setups and experiment results for five variants of proposed AV(AL) DOA model. And the baseline model is [17].

Model	Inputs			Metrics		
	$\mathbf{F}_A$	$\mathbf{X}_V$	$\mathbf{L}_V$	DOA Error(°)	Acc	F1-Score
A	√	-	-	33.48	0.32	0.38
AV(A)	√	√	-	29.91	0.47	0.63
AV(L)	√	-	√	26.64	0.35	0.90
V(AL)	-	√	√	8.72	0.84	0.90
AV(AL)	√	√	√	**7.42**	**0.86**	**0.92**
baseline	√	√	-	27.98	0.46	0.59

Table 2. The DOA Error metric of ablation experiments for different fusion methods and loss functions on the proposed AV(AL) DOA model, where 'speaker-wise' denotes the proposed BceCosLoss, and 'speaker-independent' means the BceCosLoss without multiplying the Cosine loss with the BCE loss.

Method	Add	Cat	Dot	FBP	Gated
speaker-independent	8.56°	11.6°	8.42°	9.95°	7.69°
speaker-wise	8.53°	9.84°	8.09°	9.55°	**7.42°**

a low localization accuracy. Considering the usefulness of angular vectors for the angle regression and speaker's lips for the active speaker prediction respectively, we combined them as the V(AL) model to analyze the localization performance. Significantly, the V(AL) DOA model effectively reduced the DOA Error by 19.26° absolutely relative to the baseline with a high localization accuracy. The proposed AV(AL) DOA model fused the audio-visual information and achieved the optimal performance in reducing the DOA Error by 73.48% and improving the localization accuracy by 86.95% relative to the baseline. As seen from the high F1-Score in AV(L), V(AL) and AV(AL) DOA models, the effectiveness of incorporating the lip features into the audio features and speaker angles indicated that the lip features capture the important information for active speaker prediction that is difficult to be captured by the others.

We also explored the effects on the five models by statistically calculating the DOA Errors corresponding to the number of people participating in the conversation as Fig. 2. As the number of speakers rises, the DOA Errors of the A, AV(A) and AV(L) DOA models increase sharply, which is caused by the increased overlap of utterances in multi-speaker conversation and the rapidly increasing angular range. In the V(AL) DOA and AV(AL) DOA models, the DOA Errors are controlled within a range of about 30°, demonstrating that the proposed model has high robustness in multi-speaker scenarios.

A series of ablation experiments for fusion methods were conducted. We adopted the common element-wise addition, concatenation, dot operations as compared in [25] and the multi-modal factorize bilinear pooling (FBP) [26] as the comparisons to the proposed multi-modal gated fusion. As the experimental results in Table 2 indicate, the gated fusion method achieves a relatively good performance among the other methods for better representation of the fused embedding, while the others have poorer localization abilities.

We also designed a set of experiments to compare the speaker-wise BceCosLoss with speaker-independent BceCosLoss (without multiplying Cosine loss by BCE loss) in Table 2. As seen from low DOA Error, the speaker-wise BceCosLoss has improved localization performances to a certain extent among different fusion strategies, showing its validity in jointly optimizing BCE loss and Cosine loss.

Fig. 3 is a randomly selected localization example of the baseline and our proposed AV(AL) model. The three people sitting in the room had tiny movements in conversation. As illustrated in red boxes, the missing detections of three speakers in the baseline were rectified by the AV(AL) model with the help of the speaker lips particularly in overlapping segments.

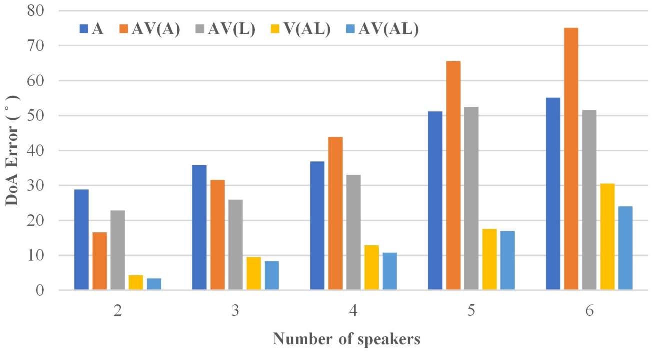

Fig. 2. A comparison of the DOA Error among five models with the increasing number of speakers involved in the conversation.

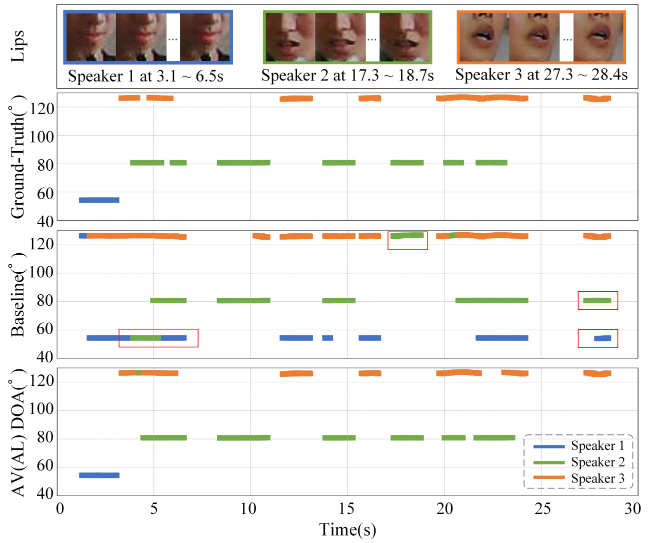

Fig. 3. An example for comparing the localization ability of the proposed AV(AL) DOA model with the baseline model.

4. CONCLUSION

In this paper, we propose a novel audio-visual multi-speaker DOA estimation model which incorporates speaker's lips by gated fusion. The proposed AV(AL) DOA model fuses speaker lip features with audio features and video angular vectors to make full use of video information. We also explore the effectiveness of a tri-modal gated fusion module with other fusion strategies, and a speaker-wise BceCosLoss to enhance the localization capability. A set of ablation experiments were conducted to validate the effectiveness and auxiliary of video lip features to DOA estimation. The proposed model achieves an excellent localization performance with a 73.48% reduction in DOA Error while improving localization accuracy by 86.95% relatively. It is inevitable that incorporating visual modality into audio modality requires greater computational complexity, but the performance of the model can be improved significantly. In the future we will explore effective techniques to compress the model and improve its real-time capability for DOA estimation applications.

5. ACKNOWLEDGE

This work was supported by the National Natural Science Foundation of China under Grant No. 62171427.

6. REFERENCES

[1] P Jeyasingh and M Mohamed Ismail, "Real-time multi source speech enhancement based on sound source separation using microphone array," in *2018 Conference on Emerging Devices and Smart Systems (ICEDSS)*. IEEE, 2018, pp. 183–187.

[2] Shengkui Zhao, Saima Ahmed, Yun Liang, Kyle Rupnow, Deming Chen, and Douglas L Jones, "A real-time 3d sound localization system with miniature microphone array for virtual reality," in *2012 7th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. IEEE, 2012, pp. 1853–1857.

[3] Xiong Xiao, Shengkui Zhao, Duc Hoang Ha Nguyen, Xionghu Zhong, Douglas L Jones, Eng-Siong Chng, and Haizhou Li, "The ntu-adsc systems for reverberation challenge 2014," in *Proc. REVERB challenge workshop*. Spoken Language Systems MIT Computer Science and Artificial Intelligence Laboratory, 2014, p. o2.

[4] 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.

[5] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[6] Jacek P. Dmochowski, Jacob Benesty, and SofiÈne Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2510–2526, 2007.

[7] Jacob Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *the Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[8] Fabio Vesperini, Paolo Vecchiotti, Emanuele Principi, Stefano Squartini, and Francesco Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2016, pp. 1–6.

[9] Harshavardhan Sundar, Weiran Wang, Ming Sun, and Chao Wang, "Raw waveform based end-to-end deep convolutional network for spatial localization of multiple acoustic sources," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4642–4646.

[10] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1462–1466.

[11] Thi Ngoc Tho Nguyen, Woon-Seng Gan, Rishabh Ranjan, and Douglas L Jones, "Robust source counting and doa estimation using spatial pseudo-spectrum and convolutional neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2626–2637, 2020.

[12] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018.

[13] Christopher Schymura, Benedikt Bönninghoff, Tsubasa Ochiai, Marc Delcroix, Keisuke Kinoshita, Tomohiro Nakatani, Shoko Araki, and Dorothea Kolossa, "PILOT: Introducing Transformers for Probabilistic Sound Event Localization," in *Proc. Interspeech 2021*, 2021, pp. 2117–2121.

[14] Bing Yang, Hong Liu, and Xiaofei Li, "Srp-dnn: Learning direct-path phase difference for multiple moving sound source localization," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 721–725.

[15] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu, "Egocentric deep multi-channel audio-visual active speaker localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10544–10552.

[16] Xinyuan Qian, Qiquan Zhang, Guohui Guan, and Wei Xue, "Deep audio-visual beamforming for speaker localization," *IEEE Signal Processing Letters*, vol. 29, pp. 1132–1136, 2022.

[17] Qing Wang, Hang Chen, Ya Jiang, Zhe Wang, Yuyang Wang, Jun Du, and Chin-Hui Lee, "Deep learning based audio-visual multi-speaker doa estimation using permutation-free loss function," in *2022 13th ISCSLP*, 2022, pp. 250–254.

[18] Hang Chen, Jun Du, Yu Hu, Li-Rong Dai, Bao-Cai Yin, and Chin-Hui Lee, "Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement," *Neural Networks*, vol. 143, pp. 171–182, 2021.

[19] John Edison Arevalo Ovalle, Thamar Solorio, Manuel Montes-y-Gómez, and Fabio A. González, "Gated multimodal units for information fusion," in *5th ICLR, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017.

[20] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic, "Lipreading using temporal convolutional networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6319–6323.

[21] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *3rd ICLR, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2015.

[22] Ilya Loshchilov and Frank Hutter, "SGDR: stochastic gradient descent with warm restarts," in *5th ICLR, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[23] Hang Chen, Jun Du, Yusheng Dai, Chin-Hui Lee, Sabato Marco Siniscalchi, Shinji Watanabe, Odette Scharenborg, Jingdong Chen, Bao-Cai Yin, and Jia Pan, "Audio-visual speech recognition in misp2021 challenge: Dataset release and deep analysis," *Group*, vol. 1, pp. 2, 2022.

[24] Harold W Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[25] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding, "Towards accurate scene text recognition with semantic reasoning networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12113–12122.

[26] Yuanyuan Zhang, Zi-Rui Wang, and Jun Du, "Deep fusion: An attention guided factorized bilinear pooling for audio-video emotion recognition," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.