Egocentric Deep Multi-Channel Audio-Visual Active Speaker Localization

Hao Jiang, Calvin Murdock, Vamsi Krishna Ithapu

Reality Labs Research at Meta
haojiang,cmurdock,ithapu@fb.com

1 Introduction

Augmented reality devices have the potential to enhance human perception and enable other assistive functionalities in complex conversational environments. Effectively capturing the audio-visual context necessary for understanding these social interactions first requires detecting and localizing the voice activities of the device wearer and the surrounding people. These tasks are challenging due to their egocentric nature: the wearer's head motion may cause motion blur, surrounding people may appear in difficult viewing angles, and there may be occlusions, visual clutter, audio noise, and bad lighting. Under these conditions, previous state-of-the-art active speaker detection methods do not give satisfactory results. Instead, we tackle the problem from a new setting using both video and multi-channel microphone array audio. We propose a novel end-to-end deep learning approach that is able to give robust voice activity detection and localization results. In contrast to previous methods, our method localizes active speakers from all possible directions on the sphere, even outside the camera's field of view, while simultaneously detecting the device wearer's own voice activity. Our experiments show that the proposed method gives superior results, can run in real time, and is robust against noise and clutter.

2 Related works

Single and multi-channel sound source detection and localization problems have classically been studied by speech and audio signal processing communities [21, 20, 11]. These approaches are sensitive to room acoustics and noisy backgrounds and may be unreliable when multiple sources are present. More recently, machine learning has been used for direction of arrival estimation with some success [12, 13, 19, 29]. Although these methods improve upon the traditional approaches, the lack of visual information limits the efficacy of these systems in real-word settings.

The computer vision community has seen a surge in audio-visual learning research, in particular due to datasets like the AVA Speech and Activity corpus [22], Voxconverse [23], and Voxceleb [24]. For action and activity recognition, several studies have shown evidence that audio disambiguates certain visually ambiguous cues [27, 28]. Audio-visual models have been explored for speech

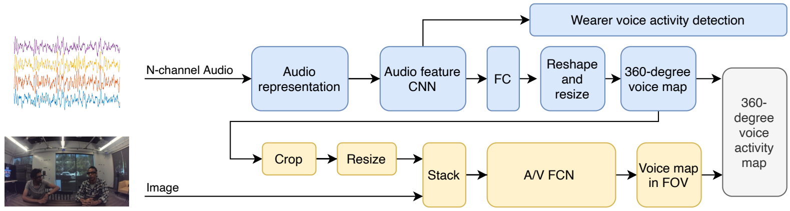

Fig. 1. Egocentric multi-channel audio-visual localization. Our end-to-end deep network detects a 360° voice activity map and the wearer's voice activity at the same time.

recognition [25], sound source detection [8–10], multiple source separation [5–7, 17], localization of sounds in a 2D image [1, 4, 30], 3D scene navigation guided by audio [26], and others.

Transformer networks have been proposed for single-channel active speaker detection [14]. More recently, turn-taking has also been studied as a means to improve detection performance [16]. A related problem is that of speech separation, which singles out a speaker's voice by using both audio and cropped facial images [5, 7, 17]. Although extensively studied, single-channel speaker detection from an egocentric perspective is still a challenging problem due to substantial device motion, occlusions, reduced visibility of speakers' faces, and noise induced by overlapping and interrupting speakers.

Single-channel audio-visual localization in exocentric settings has received much attention lately [3, 8–10, 15]. Due to the lack of multiple channels, localization is restricted to the image frame in a manner similar to traditional visual object localization. To train multi-channel AV features, a self-supervised method was proposed for face localization using audio around a target frame with a reference frame from another part of the same video as input [31]. However, a 360-degree version of this requires panoramic images and aligned audio spherical harmonics. Both of these are restrictive and not available in our AR problem setting. In [2] the authors propose an audio-visual model that can process binaural (two-channel) audio for sound source localization. However, the system cannot be extended to multi-channel settings, and is restricted to localizing targets within the visual field of view.

3 Egocentric Active Speaker Localization

Given multi-channel audio-visual data captured using AR glasses with a microphone array and RGB camera, we define the egocentric ASL problem as the detection and spatio-temporal localization of all the active speakers in the scene including the voice activity of the device wearer.

Fig. 1 illustrates the proposed egocentric ASL framework. Our method is an end-to-end deep learning model which takes the raw audio and video as input and estimates the active speaker activity heat map ($\mathbf{V}$) and wearer's voice activity

(**W**) directly. The framework has two networks: an audio network cascade ($\mathcal{A}$) and an audio-visual network cascade ($\mathcal{AV}$). $\mathcal{A}$ converts raw multi-channel audio and compacts a 2D representation aligned to each video frame, which is then used to extract relevant features using a convolutional neural network to estimate a direction of arrival estimate for the sources in the scene. $\mathcal{AV}$ then utilizes the outputs from $\mathcal{A}$ and incorporates visual information using another network. The resulting outputs from both $\mathcal{A}$ and $\mathcal{AV}$ are then combined to compute **V** and **W**.

We train the network in two stages. In the first stage, we train the audio-only and audio-visual network together without the wearer's voice activity classification network. In the second stage, we fix the audio feature layer's weights and train the fully connected network to predict the wearer's voice activity. More details of the proposed method can be found in [32].

4 Experiment Results

We evaluate our method using the EasyCom [18] dataset, a multi-channel audio-visual dataset that includes around 6 hours of egocentric videos of conversations within a simulated noisy environment. We use the RGB egocentric video together with the multi-channel audio from the four fixed microphones in our experiments. The dataset has 12 video sessions. We use sessions 1-3 for testing and the remaining 9 sessions for training.

We compare the proposed method in different variations against other active speaker detection and localization methods. The methods in the evaluation include:

`Ours AV(·)`: Variations of our method including different combinations of feature representations (`cor`: cross correlation, `eng`: energy, `spec`: spectrogram, and `box`: head bounding boxes).

`DOA+headbox`: A state-of-the-art signal processing method [20] for extracting spherical direction-of-arrival (DOA) energy maps from the 4 microphones on the glasses combined with head detection bounding boxes for active speaker detection.

`DOA+image`: A deep neural network trained to localize active speakers using both traditional signal processing DOA maps [20] and video frames as inputs.

`AV-rawaudio`: A deep neural network trained using multi-channel raw audio and video as the input.

Mouth region classifier (`MRC`): A visual-only method for classifying active speech from cropped images of mouth regions extracted from a 68-point facial key point detector.

`TalkNet` [14]: A transformer-based single-channel audio-visual active speaker detection method that gave state-of-the-art results in the AVA active speaker detection challenge. We use the method in two modes: `TalkNet(AVA)` trained on the AVA dataset and `TalkNet(EasyCom)` trained on EasyCom.

`BinauralAVLocation` [2]: A two-channel audio-visual method for sound source localization.

	ASL mAP
Ours AV(cor)	84.14
Ours AV(cor+eng)	83.32
Ours AV(cor+box)	86.25
Ours AV(cor+eng+box)	86.32
Ours AV(spec)	85.49
Ours AV (eng)	62.68
Ours AV(cor)-2ch	80.00
Ours AV(spec)-2ch	83.30
AV-rawaudio	72.32
DOA+headbox	52.62
DOA+image	54.27
MRC (AVA)	46.60
MRC(EasyCom)	64.24
TalkNet (AVA)	69.13
TalkNet (EasyCom)	44.24
BinauralAVLoc	60.75

(a)

	Mean E1	Std1	Mean E2	Std2
Ours AV (cor)	16.77	12.63	6.56	8.77
Ours AV (spec)	8.81	9.63	6.21	6.89
DOA	129.82	18.26	46.45	21.50
DOA+image	66.81	7.89	36.48	8.97
AV-rawaudio	40.14	10.55	140.75	19.58

(b)

	Wearer audio activity mAP
Ours(cor)	90.20
Ours(cor+eng)	90.13
Ours(eng)	88.89
Ours(spec)	91.69
Ours(cor)-2ch	87.66
Ours(spec)-2ch	90.14
Eng(single channel)	76.71
AV-rawaudio	87.29

(c)

Fig. 2. (a): Comparison of mAPs in the visual field of view. (b): Comparison of full 360° spherical voice activity localization errors measured in degrees. (c): Camera wearer voice activity detection. Numbers in (a) and (c) show percentages.

4.1 Within-View Active Speaker Detection

We first evaluate the mean average precision (mAP) of active speaker localization detections within the camera's field of view. As shown in Fig. 2 (a), our methods give much higher mAP than all of the competing methods.

4.2 Spherical Active Speaker Localization

One unique property of our proposed method is that it gives a full 360° spherical speaker localization result. We compare our method with methods that use traditional DOA maps and the audio-visual variation with raw audio input. As shown in Fig. 2 (b), our method gives the lowest angular errors.

4.3 Wearer Voice Activity Detection

Another unique property of the proposed method is that it can simultaneously detect the voice activity of the person wearing the recording glasses. Camera wearer audio activity detection is a new task. We construct different natural solutions in the comparison. As shown in Fig. 2 (c), our proposed method gives better results than the competing methods.

The proposed method runs in real time at over 180 frames per second using a single GTX2080Ti GPU with about 50% utilization. The proposed method also has a smaller latency compared to traditional signal processing methods, which require estimating signal statistics over longer windows of time.

References

1. P. Morgado, N. Vasconcelos, T. Langlois, O. Wang. Self-Supervised Generation of Spatial Audio for 360-degree Video, NIPS 2018.
2. X. Wu, Z. Wu, L. Ju, S. Wang. Binaural Audio-Visual Localization, AAAI-21.
3. A. Owens, A. A. Efros Audio-Visual Scene Analysis with Self-Supervised Multisensory Features, ECCV 2018.
4. A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, I. S. Kweon. Learning to Localize Sound Source in Visual Scenes, CVPR 2018.
5. A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, M. Rubinstein. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation, ACM Transactions on Graphics, Vol. 37, No. 4, pp 1-11, August 2018.
6. R. Gao, R. Feris, K. Grauman. Learning to Separate Object Sounds by Watching Unlabeled Video, CVPR 2018.
7. T. Afouras, J.S. Chung, A. Zisserman. The Conversation: Deep Audio-Visual Speech Enhancement, arXiv:1804.04121.
8. I. D. Gebru, X. Alameda-Pineda, R. Horaud, F. Forbes. Audio-visual speaker localization via weighted clustering, IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2014.
9. R. Qian, D. Hu, H. Dinkel, M. Wu, N. Xu, and W. Lin. Multiple Sound Sources Localization from Coarse to Fine, ECCV 2020.
10. H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, A. Zisserman. Localizing Visual Sounds the Hard Way, CVPR 2021.
11. C. Rascon, I. Meza. Localization of Sound Sources in Robotics: A Review. Robotics and Autonomous Systems, Volume 96, October 2017, Pages 184-210.
12. S. Adavanne, A. Politis, T. Virtanen. Direction of Arrival Estimation for Multiple Sound Sources Using Convolutional Recurrent Neural Network, European Signal Processing Conference (EUSIPCO), 2018.
13. T.N.T. Nguyen, W-S. Gan, R. Ranjan, D.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 (Volume: 28)
14. R. Tao, Z. Pan, R.K. Das, X. Qian, M.Z. Shou, and H. Li. Is Someone Speaking? Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection. The 29th ACM International Conference on Multimedia, 2021.
15. O. Kopuklu, M. Taseska, G. Rigoll. How To Design a Three-Stage Architecture for Audio-Visual Active Speaker Detection in the Wild, ICCV 2021.
16. T.-D. Truong, C. N. Duong, T. D. Vu, H. A. Pham, B. Raj, N. Le K. Luu. The Right to Talk: An Audio-Visual Transformer Approach. ICCV 2021.
17. R. Gao and K. Grauman. VisualVoice: Audio-Visual Speech Separation with Cross-Modal Consistency. CVPR 2021.
18. J. Donley, V. Tourbabin, J. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, R. Mehra. EasyCom: An Augmented Reality Dataset to Support Algorithms for Easy Communication in Noisy Environments, arXiv:2107.04174
19. P.-A. Grumiaux, S. Kitic, L. Girin, A. Guerin. A Survey of Sound Source Localization with Deep Learning Methods, arXiv:2109.03465.
20. V. Tourbabin, J. Donley, B. Rafaely, R. Mehra. Direction of Arrival Estimation in Highly Reveberant Environments Using Soft Time-Frequency Mask, 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.

21. D. P. Jarrett, E. A.P. Habets, P. A. Naylor. Theory and Applications of Spherical Microphone Array Processing, Springer Topics in Signal Processing, 9.
22. J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi, C. Pantofaru. AVA-ActiveSpeaker: An Audio-Visual Dataset for Active Speaker Detection, arXiv:1901.01342.
23. J. S. Chung, J. Huh, A. Nagrani, T. Afouras, A. Zisserman. Spot The Conversation: Speaker Diarisation in The Wild, ArXiv, 2020.
24. J. S. Chung, A. Nagrani, A. Zisserman. VoxCeleb2: Deep Speaker Recognition, INTERSPEECH, 2018.
25. T. Afouras, J. S. Chung, A. Senior, O. Vinyals, A. Zisserman. Deep Audio-visual Speech Recognition, TPAMI, December, 2018.
26. C. Chen, U. Jain, C. Schissler, S. V. Amengual Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, K. Grauman. SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020.
27. E. Kazakos, A. Nagrani, A. Zisserman, D. Damen. EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. ICCV 2019.
28. F. Xiao, Y. J. Lee, K. Grauman, J. Malik, C. Feichtenhofer. Audiovisual SlowFast Networks for Video Recognition, arXiv, 2020.
29. C. Gan, H. Zhao, P. Chen, D. Cox, A. Torralba. Self-Supervised Moving Vehicle Tracking With Stereo Sound, ICCV 2019.
30. J. Ramaswamy, S. Das. See the Sound, Hear the Pixels, WACV 2020.
31. K. Yang, B. Russell, J. Salamon. Telling Left from Right: Learning Spatial Correspondence of Sight and Sound, CVPR 2020.
32. H. Jiang, C. Murdock, V. K. Ithapu. Egocentric Deep Multi-Channel Audio-Visual Active Speaker Localization, CVPR 2022.