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

Abstract

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.

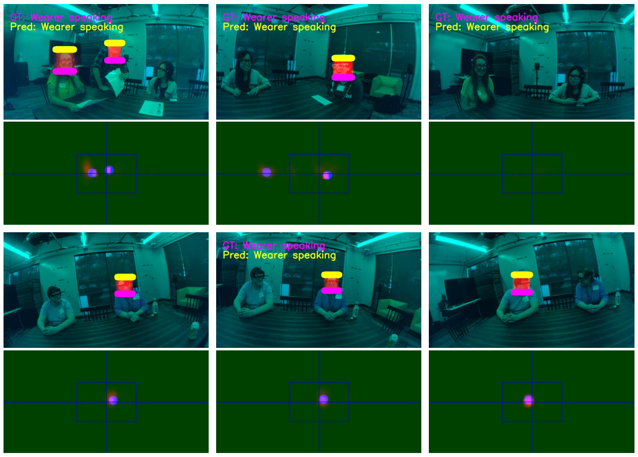

Figure 1. Our novel multi-channel audio-visual deep network localizes active speakers from any direction on the sphere, even beyond the camera's field of view. Here, predicted active speaker probability heat maps are shown in the red channels of both the images (rows 1,3) and voice maps (rows 2,4). These voice maps are 360×180 cylindrical 2D projections of the sphere, where each pixel corresponds to a direction in the device wearer's local 3D coordinate system and the camera's limited field of view is approximated by the central blue rectangle. Ground truth active speakers are shown as purple bars below head bounding boxes and as blue dots in the voice maps, while our method's thresholded predictions are shown as yellow bars. The overlaid text indicates ground truth (purple) and predicted (yellow) wearer voice activity detections.

1. Introduction

Understanding conversational context and dynamics from an egocentric perspective is vital for creating realistic and useful augmented reality (AR) experiences. These attributes characterize the interactions of multiple speakers in a given scene with the AR device wearer (i.e., *ego*). An example such device may consist of glasses with outward looking cameras and microphones so that audio-visual data is captured from the wearer's point of view. Modeling these attributes involves not only detecting and tracking people within a scene, but also localizing the voice activity within a conversation. In this work, we focus on the task of active speaker localization (ASL) with the goal of detecting the spatio-temporal location of all active speakers both within and outside the camera's field of view (FOV). Closely related to the problem of active speaker detection (ASD), ASL involves estimating the relative direction of arrival of speech from an egocentric perspective. In this paper, active speakers typically correspond to the people who are speaking and 'driving' the conversations. The elements of our proposed egocentric ASL problem are illustrated in Fig. 1.

A good ASL system needs to account for the changing orientations of speakers from an egocentric point of view and be robust to speakers moving in and out of the visual field of view. In particular, natural conversations entail significant overlap between different speakers' voice activity and involve one or more speakers interrupting each other — a classical attribute in conversational ecology called turn-taking. Such a system should also ideally be agnostic to the number of microphone channels, thereby allowing for generalization to different AR devices with varying numbers of audio and/or visual channels. Note that the device wearer may also be an active speaker during the conversation whose voice is naturally amplified due to their closeness to the device microphones. An ASL system must account for this *false* amplification that may nullify competing active speakers in the scene. In this work, we propose a real-time audio-visual ASL system that addresses these aspects to effectively localize active speakers potentially outside of the visual FOV by leveraging audio recorded from a device-mounted microphone array.

We propose a new end-to-end deep neural network trained to tackle this problem. Our network is partitioned into two branches: an audio network and an audio-visual network. The audio network builds useful representations for constructing a low-resolution sound source localization map with a full 360° FOV by utilizing spatio-temporal correlations across different channels. The audio-visual network then combines the extracted audio features with the corresponding video frames, resulting in a higher resolution activity map for the camera's FOV. Visual cues such as the person's mouth movement, facial expressions, and body pose are extracted here and combined with audio features for computing a joint representation. The final 360° active speaker map is a combination of the low-resolution audio-only map and the high-resolution audio-visual map. In addition, the device wearer voice activity detector shares the features from the audio network, and our model estimates the relative 3D orientations of the speakers in the scene from an egocentric perspective. The proposed network is also aimed at real-time applications in the immersion-driven domain of AR, enabling systems for the spatialization and localization of audio-visual activity in a world-locked frame of reference. Lastly, the lack of reliable multi-channel conversational datasets is another limiting factor for building in-the-wild ASL systems. To that end, we build and evaluate our approach using a very recent egocentric conversations dataset called EasyCom [18].

Our contributions are:

1. We propose the new problem of active speaker localization (ASL), predicting the relative locations of all active speakers in the auditory scene using egocentric multi-channel audio and video.

2. To solve this problem, we propose a real-time egocentric audio-visual system with a full 360° field of view. Our novel multi-channel audio-visual deep network can effectively learn from different audio features and microphone arrays without structure changes.

3. We evaluate our method on the EasyCom dataset and demonstrate significantly improved results in comparison to previous audio-visual ASD approaches.

1.1. Related Work

Single and multi-channel sound source detection and localization problems have classically been studied by speech and audio signal processing communities [11, 20, 21]. Most of these works are based on source separation and voice activity detection, and they mainly assume that there is one speaker in the audio stream who dominates the others (i.e., a high signal-to-noise ratio). The primary characteristic of these methods is to build auto-correlation and cross-correlation functions across different channels to account for timing and level differences caused by microphone placement. However, 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. Furthermore, most multi-channel approaches assume fixed, stationary microphone arrays, which may lead to poor performance with moving arrays in egocentric 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]. These approaches are driven by building correspondences between audio and visual modalities, thereby resulting in robust joint representations that improve upon their audio-only or image-only counterparts. 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 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.

A bulk of the audio-visual learning models follow a simple recipe: audio inputs are converted to spectrogram images which are then jointly processed with video frames. In addition to traditional network architectures, transformer networks have also 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]. The voice energy of the enhanced speech can then be used to detect active speakers. 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. Most current methods also induce significant latency in detection, which would be ineffective for real-time AR experiences.

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. These methods either utilize audio-visual joint embeddings similar to those in active speaker detection, or they train audio-visual joint classification modules as the backbone for modality fusion. 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.

2. 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. Let $\mathbf{A}_i$ with $(i = 1..N)$ denote the audio signals captured via N-channel microphone array and $\mathbf{I}$ denote the video from the RGB camera. The audio signals are normalized to the range [-1,1] based on the maximum bit length of audio samples. At each time instant t, given a segment of audio $\mathbf{A}_i^t$ and the corresponding video frame $\mathbf{I}^t$, we estimate two outputs: a heat map $\mathbf{V}_{\alpha,\beta}^t$ of activity in the scene and the device wearer activity $\mathbf{W}$. $\mathbf{V}_{\alpha,\beta}^t$ is a 2D matrix where each element gives the probability of a sound source being present at particular relative angles (α, β) at the time instant t, where $\alpha \in [-180, 180]$ and $\beta \in [-90, 90]$ correspond to azimuth (horizontal) elevation (vertical) respectively. Although we focus on human speech in this work, the proposed framework is applicable to any sounds of interest.

Fig. 2 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 ($\mathbf{W}$) 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 $\mathbf{V}$ and $\mathbf{W}$.

2.1. Audio Representation

In this paper, we consider three audio representations and design our deep network so that it can take these different representations together with video as input in the same fashion. Our experiments show these audio representations are stronger than the raw audio. These different audio representations have different properties that are suitable for different use cases.

Our first audio representation is adapted from the complex spectrogram representation [2]. For audio with sampling rate of $48kHz$ and video frame rate at $20Hz$, we compute the short-time Fourier transform (STFT) and extract 100 discrete Fourier transforms (DFTs) of length 200 to align with each video frame. The real and imaginary parts of the DFTs from all the channels are stacked together along the depth axis to form the multi-channel 2D tensor.

In addition, we further propose a 2D audio representation that captures the cross correlation between all pairs of the audio channels. Unlike spectrograms, this representation is mostly speaker invariant. Assuming the audio sample n matches the time stamp of video frame at time t, the cross correlation between channels p and q is defined as

$$C_{p,q}(n,m) = \frac{\sum_{k=0}^{K}[A_p(n-k)A_q(n-k+m)]}{\sqrt{\sum_{k=0}^{K} A_p(n-k)^2)}\sqrt{\sum_{k=0}^{K} A_q(n-k+m)^2)}},$$

where $m \in [-L, L]$, and K and L are hyperparameters. In our experiments, audio signals have sampling rate $48kHz$, $K = 1200$ and $L = 50$. In a discrete format, $C_{p,q}(n,m)$ is a vector of length $2L + 1$ at each time n that characterizes the time shifts of different audio channels due to different paths of the sound transmission along with other fine-grained channel couplings. From this, we construct a 2D audio representation at each time n, which is a stack of all the vectors $C_{p,q}(n,m)$ for each (p,q) pair.

The short-time energy of audio is a feature that is invariant to sound sources and easy to compute. Therefore, we also include a separate measure of the energies from each audio channel, $E_p(n) = (\sum_{k=0}^{K} A_p(n-k)^2)^{0.5}$. Using this, we stack $\mathbf{e_p(n)}$ for each p, where $\mathbf{e_p(n)}$ is a vector that duplicates the $E_p(n)$ by $2L + 1$ times, to form a 2D energy map. These features can also be combined to form richer representations. Fig. 3 illustrates how the combined cross correlation and energy feature correspond to the audio events in videos. The cross-correlation, energy and the combined 2D feature are further resized. In this paper, the width and height are resized to 128.

2.2. Audio Activity Network

The audio activity network predicts a rough 360° audio activity map and the voice activity of the device wearer. Its structure is shown in Fig. 4. The feature extraction network is adapted from the first several layers of a ResNet18 network whose coefficients are pre-trained on ImageNet.

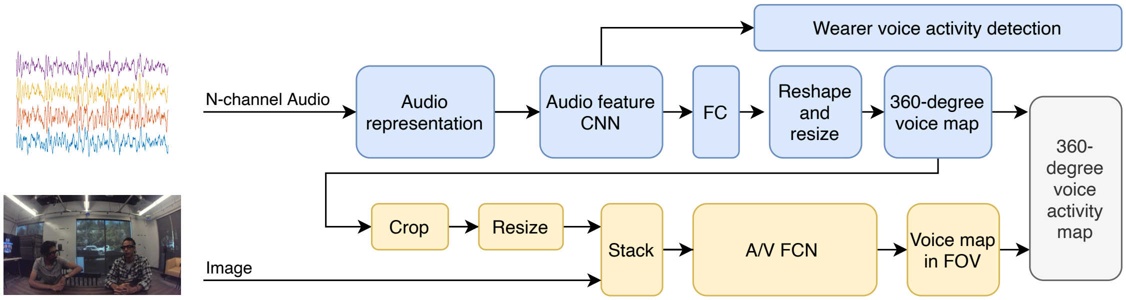

Figure 2. Egocentric multi-channel audio-visual localization. Our end-to-end deep network detects a 360° voice activity map and the wearer's voice activity at the same time.

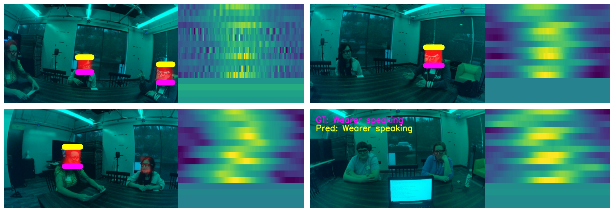

Figure 3. Odd columns: video frames overlaid with voice activity labels. Even columns: vertical stack of the audio cross correlation and energy feature maps.

The first convolutional layer is modified to match the channel number of different audio representations. The feature extraction network maps the audio 2D representation to a compact feature that quantifies the spatial and voice characteristics of audio signals in the scene. The extracted features are flattened and passed to two fully connected layers, which are further reshaped to two 90×45 maps. The two maps are stacked and resized to a 180×90 one-hot representation half the size of the full 360° audio activity map. This network thus predicts the voice activity probability from each direction with an angular resolution of 2°.

One important design here is to generate the one-hot representation of the heat map and train using cross-entropy loss. This gives more stable results than directly regressing a single heat map of the audio activity using L1 or L2 losses. A pixel-level regression network would have a larger search space due to increased degrees-of-freedom leading to training instability.

The audio activity map is also used to simultaneously estimate the wearer's voice activity. Due to the spatial position of the wearer's mouth relative to the microphones and the loudness of the wearer's voice, the 2D feature representation learned by the audio localization network also provides useful information for detecting whether the device wearer is speaking. To accomplish this, the audio feature extraction is shared with the 360° audio map prediction, and wearer voice activity detection is performed by a separate head that consists of two fully-connected layers trained to predict probability with a cross-entropy loss.

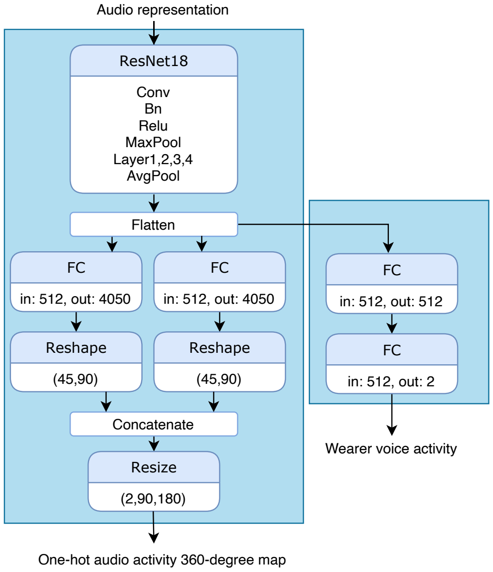

Figure 4. The audio activity network.

2.3. Audio-Visual Network

With only multi-channel audio available for speaker localization, the spatial resolution is low. This is due to the inherent physics of sound propagation and the limitations of compact microphone arrays. We therefore also take advantage of video frames to further improve the estimation result. Images not only increase spatial resolution, but also provide extra informative cues related to voice activity, such as mouth movement, facial expression, and hand gestures.

In this paper, we propose a different approach to fusing audio and visual information from previous audio-visual methods: we directly stack the video frames with the estimated voice activity map from the audio network. Since the rough 360° voice map from the audio network is defined on the unit sphere and the grids are horizontal and vertical angles, we need a procedure to align the audio map to the corresponding video frames. Even though we can map each grid in the voice map to the image, we find a simpler crop-

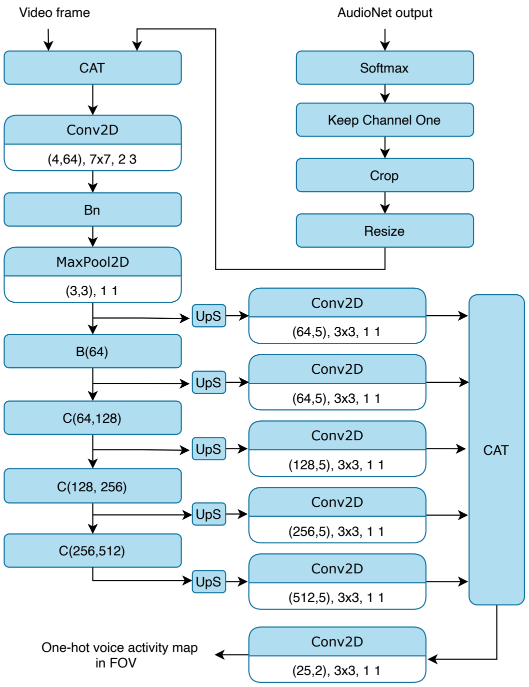

Figure 5. Audio-visual network. The blocks $B(p)$ and $C(p, q)$ are defined in Fig. 6. For 2D convolution layers, the parameters are input channel number, output channel number, convolution kernel size, stride and padding. For maxpool layer, the parameters are pooling kernel size, stride and padding.

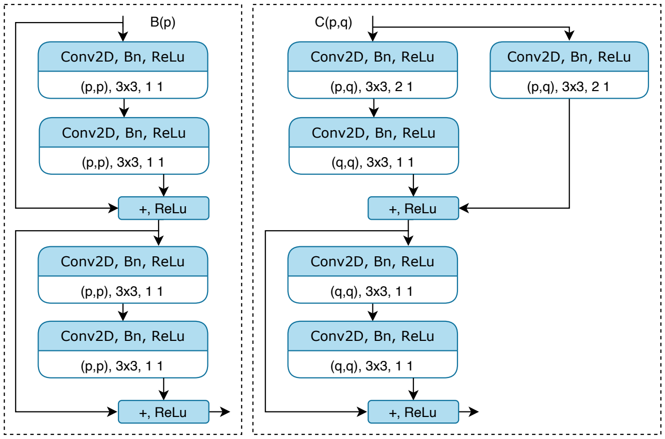

Figure 6. Residual blocks in the audio-visual network.

ping and scaling method is sufficient due to the low resolution of the audio map. More specifically, we crop the region from the audio map within the horizontal and vertical angles corresponding to the four corners of the image. The scaling procedure then upsamples the region so that the audio map in the FOV is aligned with the input video. These operations are integrated in the audio-visual network. As shown in Fig. 5, the fused audio map and the corresponding color video frame form a tensor with depth of 4, which is sent to a fully-convolutional network to estimate the refined voice activity map in the camera's field of view. In this paper, the video resolution is 640×360.

With such a design, if the faces are visible, the audio-visual network is able to take advantage of image features such as the appearance of the mouth and facial expression to localize audio activity. Due to its wide effective receptive field, the proposed network can also learn to extract other visual features such as body pose. Unlike previous methods, if the faces are not visible, our proposed method can still function because the audio activity map gives the locations of the potential speakers in the scene.

We combine the rough 360° heat map and the more detailed heat map in the FOV. In this paper, we simply double the rough 360° heat map outside of the FOV and add the refined heat map and the rough 360° heat map inside the FOV to generate the final estimation.

2.4. Model Training

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.

The 360° voice map and the voice map in the FOV are represented differently in the ground truth. The 360° voice map is a 180×90 2D map. If there is a speaker located at (α, β), the ground truth voice map has a solid disk with radius 5 centered at the point. Such labeling is uniform for regions inside and outside of the field of view. In contrast, the voice map in the FOV has the same size as the video frames, and the active speaker in the field of view is labeled as a solid rectangle that covers the speaker's head. Therefore inside the FOV, the detection also has an attribute of size which is related to the depth of the target. The training losses are defined as follows.

The first and second stage loss functions are defined as

$$\mathcal{L}_a = \mathcal{H}(y_a, \hat{y}_{360}) + \mathcal{H}(y_{av}, \hat{y}_{fov}), \quad \mathcal{L}_b = \mathcal{H}(y_w, \hat{y}_w)$$

where $\mathcal{H}$ is the mean cross entropy, y_a and y_{av} are the one-hot output representations of the audio-only and audio-visual networks, $\hat{y}_{360}$ and $\hat{y}_{fov}$ are their corresponding ground truth audio maps, y_w is the wearer speech activity prediction, and $\hat{y}_w$ is its ground truth label. The training procedure generally converges quickly within 5 epochs.

3. Experiment Results

In this section, we evaluate the proposed method on real videos and compare it with different audio-visual approaches for active speaker detection and wearer voice activity detection. Since we consider a novel egocentric problem setting, there are no previous audio-visual methods that are directly applicable. For comparison, we adapt our multi-channel audio and video inputs to other approaches to similar problems. We also compare variations of the proposed method to justify our design choice.

3.1. Evaluation Dataset

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. The dataset is recorded using a microphone array and a RGB camera mounted on a pair of glasses. EasyCom is a challenging dataset with significant background noise, fast head motion, and motion blur. Participants may sit or walk around in the scene, and their faces and mouths are not always visible due to occlusions.

There are six microphones used for recording: four fixed to the glasses and two placed within the ears of the participants. In this paper, 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, each of which is about 30 minutes long with 4-6 participants including the camera wearer. We use sessions 1-3 for testing and the remaining 9 sessions for training. For fair comparison, we report the best numbers for all competing models trained until convergence after a sufficiently large number of epochs.

3.2. Evaluation Methods

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). In the variation that uses head bounding boxes, we set the background color outside of the detected head regions to black. We also evaluate the audio-only and video-only versions of our method in which the video or audio branches are removed from our full model.

`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. This DOA estimation method was designed to achieve more robust results in highly reverberant settings compared to previous signal processing audio localization methods. To detect active speakers in the field of view, we pool regions of the DOA map corresponding to directions within the detected head bounding boxes. If the DOA map accurately estimates sound arrival directions, then the head bounding boxes corresponding to active speakers will include higher energy values.

`DOA+image`: A deep neural network trained to localize active speakers using both traditional signal processing DOA maps [20] and video frames as inputs. The network is fully convolutional and has the same structure as the audio-visual network in our method.

`AV-rawaudio`: A deep neural network trained using multi-channel raw audio and video as the input. Aside from extracting audio features with 1D convolution layers, the overall network architecture is the same as our approach.

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. Such a scheme has been commonly used in active speaker detection. A ResNet18 network is trained to classify the cropped mouth images. We test two cases: `MRC(AVA)` trained using the AVA active speaker detection dataset [22], and `MRC(EasyCom)` only trained on EasyCom.

`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. Since this method cannot be easily extended to settings with more than two asymmetric microphones, we use only the audio channels from the two frontal microphones in our comparisons.

3.3. Within-View Active Speaker Detection (ASD)

We first evaluate the mean average precision (mAP) of active speaker localization detections within the camera's field of view. We compare against multi-channel as well as one- and two-channel audio-visual methods and visual-only method. The mAP is computed based on the scores within the ground truth head bounding boxes in each video frame. For our methods and the competing methods `DOA+headbox`, `DOA+image`, `AV-rawaudio`, and `BinauralAVLocation` we extract the voice heat map's maximum value in each ground truth head bounding box and use it as the detection score. The `MRC` and the `TalkNet` methods use the classification probability of the corresponding head box as the detection score. Both `MRC` and `TalkNet` use the ground truth head bounding boxes for testing.

As shown in Table 1, our methods give much higher mAP than all of the competing methods. Fig. 7 shows qualitative comparison results. Due to the difficulty in learning useful features from raw audio, `AV-rawaudio` gives inferior results in comparison to spectrogram and cross-correlation audio features. Background noise also causes traditional audio-only signal processing approaches to give blurry DOA maps and inaccurate target localization results. The `DOA+image` deep learning method that combines this DOA map with video frames improves performance, but still gives lower mAP than our proposed

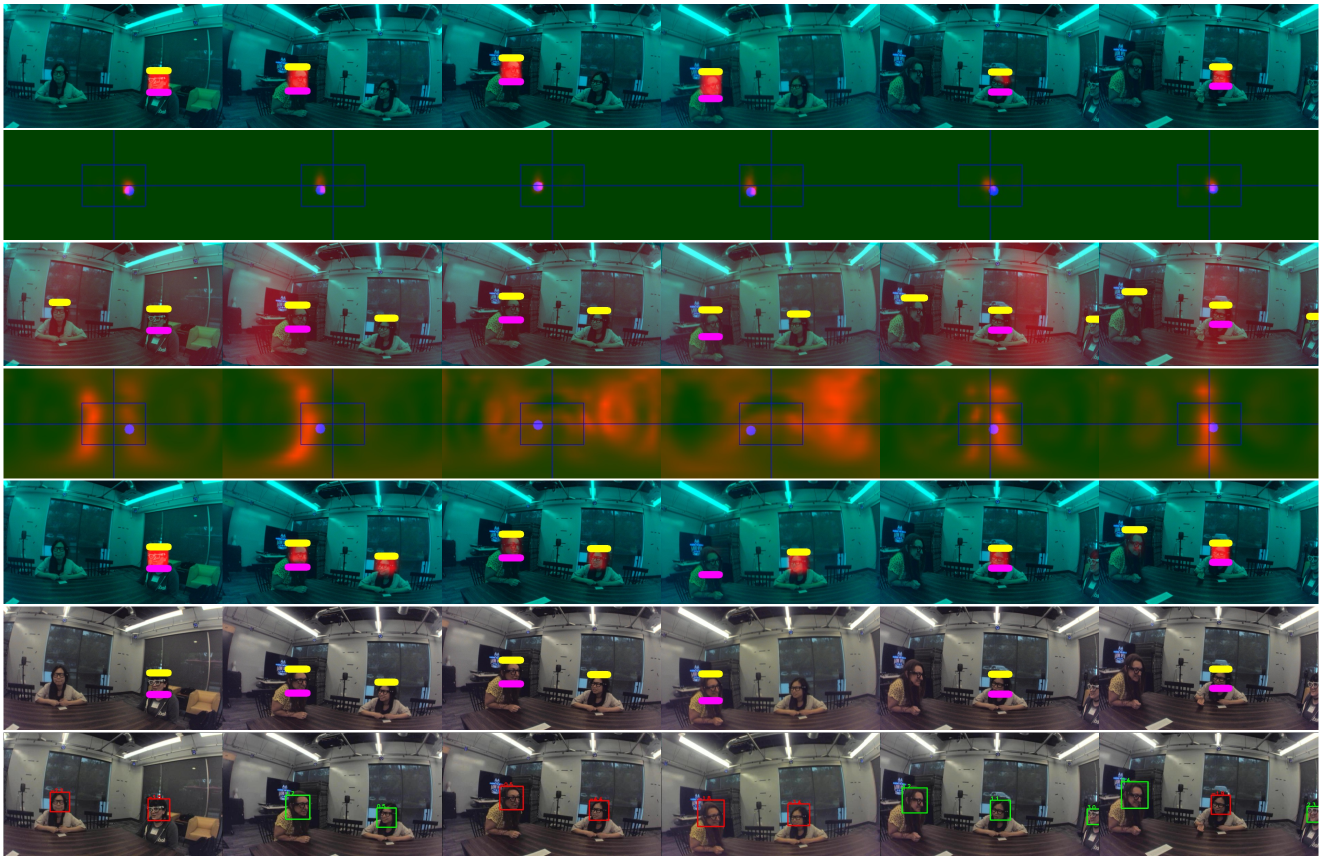

Figure 7. Qualitative comparison results. The yellow bar indicates when a person is predicted to be talking while the purple bar is the corresponding ground truth. Rows 2, 4: the predicted 360° voice map compared against the the ground truth in blue channel. Rows 1, 2: The result of `Ours AV(corr)`. Rows 3, 4: `DOA+headbox`, Row 5: `DOA+image`, Row 6: `MRC(EasyCom)`, Row 7: `TalkNet(EasyCom)`. In Row 7, green boxes indicate active speech while red boxes are inactive.

method. This emphasizes the benefit of learning spatial audio-visual representations end-to-end. Our method also gives much higher mAP than the previous video-only `MRC` and single-channel audio-visual active speaker detection method `TalkNet` trained on both the AVA dataset [22] and the EasyCom dataset. Our method greatly outperforms the `BinauralAVLocation` in both the 4-channel and 2-channel audio settings.

For different variations of the proposed method, as shown in Table 1, the energy feature is significantly worse than the other two features, while spectrogram features give slightly better mAP. The cross correlation and energy features are still attractive due to their speaker-invariant properties and thus have potential to generalize better in real applications. The cross correlation feature is also invariant to the microphone gain settings; this makes it useful when the gains need to change dynamically for best signal-noise ratio.

We also compare our audio-only and video-only variations with the full audio-visual model. In comparison to our full audio-visual method `Ours AV(cor+mag+box)` with a mAP of 86.32%, the video-only variation gave a much lower mAP of 58.44% and the audio-only version also gave a lower mAP of 78.08%. The results of `Ours AV(corr+box)` and `Ours AV(corr+eng+box)` also show that our proposed method can generalize to different environments by removing background visual information outside of head detections, which can potentially improve the result. Even with only two audio channels, our network still gave strong results that outperformed the `BinauralAVLoc` network architecture designed to leverage the symmetry of binaural audio.

3.4. Spherical Active Speaker Localization (ASL)

One unique property of our proposed method is that it gives a full 360° spherical speaker localization result. Since there is no head bounding box outside of the field of view, we use the angular error to measure the localization quality.

The metric is defined as follows: We first extract the detected target locations in the predicted voice heat map using non-maximum suppression. Every peak in the heat map with value greater than a threshold is a potential target. In

	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

Table 1. Comparison of mAPs in the visual field of view. Most of these tests use 4-channel audio, except for `Ours AV(cor)-2ch`, `Ours AV(spec)-2ch`, `BinauralAVLoc`, which use 2-channel audio, `TalkNet` which uses single-channel audio, and video-only `MRC`. Numbers show percentages.

the experiments, we set the threshold to 0. The positions in the heat map indicate the angles of directions. We compute the minimum distances from the detected points to the ground truth points in the voice heat map, whose mean is denoted as E1. We compute mean E1 and its standard deviation Std1. The corresponding metrics from the ground truth point set to the detected point set are mean E2 and Std2. We compute the distance metric in two directions to take into account both missing detections and false alarms.

Since not all the competing methods can give full 360° spherical localization results, we compare our method with methods that use traditional DOA maps and the audio-visual variation with raw audio input. As shown in Table 2, our method gives the lowest angular errors.

	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

Table 2. Comparison of full 360° spherical voice activity localization errors measured in degrees.

3.5. Wearer Voice Activity Detection (VAD)

Another unique property of the proposed method is that it can simultaneously detect the voice activity of the person wearing the recording glasses. Our method shares the learned audio features for both tasks. During the training of the camera wearer voice networks, the shared feature design freezes the network feature extraction parameters while only training the last two fully connected layers.

	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

Table 3. Camera wearer voice activity detection. `Eng(single channel)` is the naive approach of using short-time energy for wearer voice classification. Numbers show percentages.

Camera wearer audio activity detection is a new task. We construct different natural solutions in the comparison. Table 3 summarizes the comparison result. As shown in Table 3, our proposed method gives better results than the competing methods. The shared feature design in fact also gives better result than training a separate wearer voice classification model. For instance, our method using cross correlation input features gives 90.2% mAP, but if we retrain a separate wearer classifier the mAP is 88.01%. This is likely because of the additional supervision in training the localization task to explicitly suppress the wearer's speech.

Comparing to traditional signal processing approaches, our method requires more computationally expensive GPU operations. However, the proposed method is still efficient. It runs in real time at over 180 frames per second using a single GTX2080Ti GPU with about 50% utilization. More optimization could also further improve the efficiency of the network. The proposed method also has a smaller latency compared to traditional signal processing methods, which require estimating signal statistics over longer windows of time. While we only use 4 microphones in our experiments, the proposed method could be easily extended to devices with any number of microphones in any array configuration. With a larger microphone array, the proposed method has the potential to achieve even better results.

4. Conclusion

We proposed a novel multi-channel audio-visual method to tackle the 360° spherical active speaker detection problem for localizing active speakers both within and beyond an egocentric camera's visual field of view while also simultaneously predicting the wearer's voice activity. Our experiments showed that the proposed method gives superior results to competing methods and runs in real time with short latency. It has potential to enable many useful AR applications.

References

[1] P. Morgado, N. Vasconcelos, T. Langlois, O. Wang. Self-Supervised Generation of Spatial Audio for 360-degree Video, NIPS 2018. 2

[2] X. Wu, Z. Wu, L. Ju, S. Wang. Binaural Audio-Visual Localization, AAAI-21. 3, 6

[3] A. Owens, A. A. Efros Audio-Visual Scene Analysis with Self-Supervised Multisensory Features, ECCV 2018. 3

[4] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, I. S. Kweon. Learning to Localize Sound Source in Visual Scenes, CVPR 2018. 2

[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. 2

[6] R. Gao, R. Feris, K. Grauman. Learning to Separate Object Sounds by Watching Unlabeled Video, CVPR 2018. 2

[7] T. Afouras, J.S. Chung, A. Zisserman. The Conversation: Deep Audio-Visual Speech Enhancement, arXiv:1804.04121. 2

[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. 2, 3

[9] R. Qian, D. Hu, H. Dinkel, M. Wu, N. Xu, and W. Lin. Multiple Sound Sources Localization from Coarse to Fine, ECCV 2020. 2, 3

[10] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, A. Zisserman. Localizing Visual Sounds the Hard Way, CVPR 2021. 2, 3

[11] C. Rascon, I. Meza. Localization of Sound Sources in Robotics: A Review. Robotics and Autonomous Systems, Volume 96, October 2017, Pages 184-210. 2

[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. 2

[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) 2

[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. 2, 6

[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. 3

[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. 2

[17] R. Gao and K. Grauman. VisualVoice: Audio-Visual Speech Separation with Cross-Modal Consistency. CVPR 2021. 2

[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 2, 6

[19] P.-A. Grumiaux, S. Kitic, L. Girin, A. Guerin. A Survey of Sound Source Localization with Deep Learning Methods, arXiv:2109.03465. 2

[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. 2, 6

[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. 2

[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. 2, 6, 7

[23] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, A. Zisserman. Spot The Conversation: Speaker Diarisation in The Wild, ArXiv, 2020. 2

[24] J. S. Chung, A. Nagrani, A. Zisserman. VoxCeleb2: Deep Speaker Recognition, INTERSPEECH, 2018. 2

[25] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, A. Zisserman. Deep Audio-visual Speech Recognition, TPAMI, December, 2018. 2

[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. 2

[27] E. Kazakos, A. Nagrani, A. Zisserman, D. Damen. EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. ICCV 2019. 2

[28] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, C. Feichtenhofer. Audiovisual SlowFast Networks for Video Recognition, arXiv, 2020. 2

[29] C. Gan, H. Zhao, P. Chen, D. Cox, A. Torralba. Self-Supervised Moving Vehicle Tracking With Stereo Sound, ICCV 2019. 2

[30] J. Ramaswamy, S. Das. See the Sound, Hear the Pixels, WACV 2020. 2

[31] K. Yang, B. Russell, J. Salamon. Telling Left from Right: Learning Spatial Correspondence of Sight and Sound, CVPR 2020. 3