AUDIO-VISUAL TALKER LOCALIZATION IN VIDEO FOR SPATIAL SOUND REPRODUCTION

Davide Berghi and Philip J. B. Jackson

Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, UK
{davide.berghi, p.jackson}@surrey.ac.uk

ABSTRACT

Object-based audio production requires the positional metadata to be defined for each point-source object, including the key elements in the foreground of the sound scene. In many media production use cases, both cameras and microphones are employed to make recordings, and the human voice is often a key element. In this research, we detect and locate the active speaker in the video, facilitating the automatic extraction of the positional metadata of the talker relative to the camera's reference frame. With the integration of the visual modality, this study expands upon our previous investigation focused solely on audio-based active speaker detection and localization. Our experiments compare conventional audio-visual approaches for active speaker detection that leverage monaural audio, our previous audio-only method that leverages multichannel recordings from a microphone array, and a novel audio-visual approach integrating vision and multichannel audio. We found the role of the two modalities to complement each other. Multichannel audio, overcoming the problem of visual occlusions, provides a double-digit reduction in detection error compared to audio-visual methods with single-channel audio. The combination of multichannel audio and vision further enhances spatial accuracy, leading to a four-percentage point increase in F1 score on the Tragic Talkers dataset. Future investigations will assess the robustness of the model in noisy and highly reverberant environments, as well as tackle the problem of off-screen speakers.

1. INTRODUCTION

In 3D audio-visual production, meticulous attention is required to accurately align sound sources with the visual events they accompany. To produce and author an immersive experience, audio sources are generally treated as objects and manually placed in the virtual space by the producer. This approach to production is often referred to as object-based media (OBM) production [1, 2]. In OBM, each object, spanning audio, video, graphics, text, or other forms of media, is accompanied by its metadata. The metadata associated with an individual object describes specific attributes or desired behaviors of the object, such as its content or position in space over time. OBM is valued for its adaptability and interactivity, allowing for tailored experiences based on user preferences or device configuration. For example, in the context of producing immersive spatial audio, object-based audio can theoretically be rendered on any loudspeaker configuration, unlike traditional channel-based mixes, which lack such flexibility.

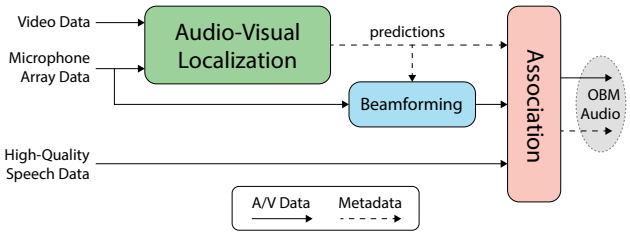

Figure 1: Pipeline for speech signals objectification proposed by Mohd Izhar *et al*. [3] and Schweiger *et al*. [4]. Positional metadata are automatically predicted by leveraging video and microphone array data. These predictions are not only used as final positional information for the spatialization of the objects but also to drive a spatial beamformer. Filtered signals extracted with the beamformer are associated with, and replaced by, the high-quality speech data recorded with Lavalier microphones. This paper focuses on the audio-visual prediction of the speaker's positional metadata.

However, while it is relatively trivial to manually spatialize synthetic audio-visual assets, this is not true for real scenes. Ideally, the ultimate spatialization should truthfully reflect the positioning and movements of the events. An example of a common audio-visual event is a moving talker. To tackle this problem, Mohd Izhar *et al*. [3] and Schweiger *et al*. [4] proposed the speaker objectification pipeline depicted in Figure 1. Automated predictions regarding the locations of active speakers are generated using an audio-visual tracker, which utilizes both video data and audio captured through a microphone array. These predictions drive the spatial listening directivity of a beamformer employed to filter the signals from the microphone array. The resulting filtered speech signals are then associated with, and substituted by, the high-quality speech recordings obtained through Lavalier microphones.

This paper addresses the automatic extraction of the speaker positional metadata, corresponding to the green block of the pipeline presented in Figure 1. Specifically, we focus on video-based horizontal active speaker detection and localization (ASDL) as videos are a widely consumed form of media, and subjective studies suggest that humans tend to be more spatially sensitive across the azimuth direction than elevation [5]. The audio-visual localization is performed on the Tragic Talkers dataset [6] and it expands upon our previous audio-only studies [7, 8]. Our previous works highlighted the benefits of employing multichannel audio captured

with a microphone array, outperforming traditional video-based audio-visual approaches for active speaker detection that employ monaural audio and rely on visual face detection. Multichannel audio provides a higher detection accuracy as it does not suffer from visual occlusions. This paper demonstrates that partnering multichannel audio with the visual modality will improve spatial accuracy while preserving the high detection rate enabled by multichannel audio.

The remainder of this document provides an overview of the related work on video-based active speaker detection and localization, describes the proposed method and the audio-visual network architecture, presents and discusses our experimental results, concludes the paper and suggests future directions for the field.

2. RELATED WORK

ASDL can be addressed through two distinct phases. Initially, the localization subtask is undertaken, wherein a visual face detector is utilized to pre-select a set of candidate speakers. Subsequently, the detected faces undergo classification into active or inactive. In computer vision, this second classification process is usually referred to as Active Speaker Detection (ASD) [9, 10, 11, 12, 13]. Researchers usually partner the video stream with the respective (mono) audio signal. Pioneering the work on video-based active speaker detection, Cutler *et al*. [14] proposed to observe the correlation between mouth motion and audio data. Haider *et al*. [15] combined lip tracking and voice activity detection (VAD) to predict who is speaking in multiparty dialogue videos. Chakravarty *et al*. [16] proposed to also include head and upper-body motion as additional visual cues to detect the active speaker. They adopted a self-supervised solution to perform ASD by training a visual network under the supervision of its audio counterpart. Subsequent works [13, 17], adopted a two-step audio-visual co-training for speaker detection and identification. They exploited the co-occurrence of speech and faces in videos to associate clusters generated from speech features with clusters generated from facial features. With the advance of deep learning techniques and the availability of larger datasets [18, 19], different yet related audio-visual tasks have been introduced, such as audio-visual lip reading [19, 20], lip-voice synchronization [21], and audio-visual speaker separation [22]. What is probably the first, large, annotated dataset for ASD was released for the ActivityNet Challenge (Task B) at CVPR 2019: the AVA-ActiveSpeaker dataset [9]. It provides 38.5 hours of audio-visual face tracks (sequences of consecutive face crops) labeled for speech activity. Since the face tracks are provided, the challenge task consists of classifying each face as active or inactive, leveraging audio and video signals. At the 2019 challenge, the first [23] and second [24] positions were achieved by leveraging 3D convolutional neural networks (CNNs). After that, Alcázar *et al.* proposed a model called Active Speaker in Context (ASC) [12]. Instead of compute-intensive 3D convolutions or large-scale audio-visual pre-training, ASC leverages context: in assessing the speech activity of a candidate speaker, it looks at any other available faces. Zhang *et al.* [10] also tackled the ASD task by leveraging contextual information and proposed the UniCon network. Tao *et al.* introduced TalkNet [11], an ASD model that leverages short- and long-term features. Additionally, motivated by the call for an ASD system that works properly outside the AVA-ActiveSpeaker dataset domain, they formed a second ASD dataset based on LRS3 [25] and VoxCeleb2 [26] called TalkSet [11]. Recently, Alcázar *et al.* [27] proposed an end-to-end ASD that unifies audio-visual feature extraction and spatio-temporal context aggregation.

However, these ASD solutions, focusing solely on the audio-visual classification of the provided face tracks, depend on visual face detection for the localization subtask. In practice, the speaker can be occluded or facing away from the camera, leading to face detection failures and subsequent degradation of the overall system performance. When this happens, monaural audio is insufficient to compensate for the visual failure, as it lacks the necessary spatial cues required to locate audio sources. In other words, the active speaker is detected only when visible. In previous studies [8, 7], we overcame this problem by leveraging multichannel audio to simultaneously address the detection and localization aspects of ASDL. Our model was able to locate the active speaker in the video frames solely from audio inputs, achieving better detection and recall rates.

Some other recent studies partnered multichannel audio with vision for audio-visual ASDL. For example, Qian *et al.* in [28] and [29] employed visual feature vectors encoding face bounding box coordinates to improve spatial accuracy. The detected face bounding boxes are represented as horizontal and vertical Gaussian-like vectors. Similarly, Wu *et al*. [30] adopted Gaussian-like vectors as visual input features and a transformer-based network [31] for their predictions. Zhao *et al*. [32] proposed a self-supervised student-teacher knowledge distillation approach to train a multichannel audio network from visual supervision. In a different study, they explored the audio-visual speaker localization from egocentric views [33]. That is, the prediction is performed from the point of view of the device "wearer", who is free to move in space and the sensors are therefore not stationary. Research groups from Meta Reality Labs released the EasyCom dataset [34]. It consists of a set of videos captured from custom AR glasses with a camera and a microphone array integrated. Jiang *et al.* [35] employed EasyCom to perform audio-visual speaker localization in an egocentric setting. Concurrently, Gurvich *et al*. [36] performed the real-time ASDL with a microphone array in 360° videos. Similarly, in this paper, we integrate multichannel audio and visual data. However, instead of leveraging visual features such as Gaussian-like vectors, we focus on the integration of audio and visual embeddings extracted with pre-trained audio and visual encoders [37].

3. METHOD

The proposed method extracts audio and visual feature embeddings with an audio and a visual encoder, respectively. The embeddings are then concatenated and fed to an attention-based unit. The output of the attention-based unit represents a joint latent audio-visual representation of the input signals, which is used to generate the final prediction through a feed-forward network. A schematic representation of the proposed model is presented in Figure 3.

3.1. Dataset

To train and evaluate the proposed method, we employed the Tragic Talkers dataset [6]. It offers sequences captured by two Audio-Visual Array (AVA) Rigs. Each AVA Rig is a light-field and sound-field sensing platform consisting of a 16-element microphone array and 11 cameras fixed on a flat perspex baffle as depicted in Figure 2. Therefore, each sensor is located in a fixed relative position and orientation with respect to the other sensors. The microphone array has a horizontal aperture of 450 mm and a vertical aperture

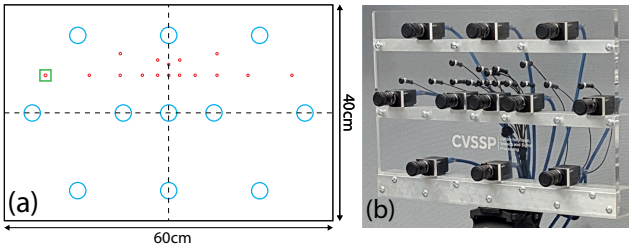

Figure 2: (a) Schematic of camera (blue circles) and microphone (red dots) positions on the AVA Rig. The green square highlights the reference microphone. (b) Photo of an AVA Rig.

of only 40 mm. This results in higher resolution when localizing audio sources along the azimuth direction than elevation, which is consistent with human perception [5]. Horizontally, the microphones are log-spaced for broad frequency coverage from 500 Hz to 8 kHz, to better support the horizontal speech band resolution. Tragic Talkers was captured in an acoustically treated studio with an average reverberation time of 0.3s in the mid 0.5-2 kHz frequency range and minimal background noise floor (SNR $\geq$ 30 dB). The dataset does not contain sequences in which the speakers talk simultaneously, off-screen talkers, or external sources of sound other than speech, making it ideal for audio-visual speaker diarization applications too. Tragic Talkers was specifically designed for OBM production research. Its content allows the implementation of the speaker objectification pipeline presented in Figure 1 [3, 4]. In fact, it provides high-quality speech signals recorded with Lavalier microphones. Additionally, the scenes are captured against a blue background that facilitates the actors' silhouette extraction, empowering the producer to extract and position the actors as desired while retaining the natural motion of the performance.

The method for ASDL proposed in this paper necessitates a single video feed with the microphone array. Leveraging the multiple views available enables us to extend the network training by choosing the relevant camera perspective. Consequently, a one-hot vector denoting the selected view is appended to the input data, augmenting the dataset with diverse camera perspectives and enabling the network to learn the correct mapping to the desired viewpoint. The sequences include one or two actors positioned at a distance of about 3–4 m, engaging in monologues, conversations, and interactive scenes involving movement and occlusion. Tragic Talkers comprises 30 scenes captured with two AVA Rigs. Each rig's audio-visual stream is employed independently, i.e., 16-channel audio is used to predict the speaker's position within any of the 11 camera perspectives of the rig. So the dataset's 30 scenes provide 60 rig sequences, each offering 11 viewpoints. The dataset is partitioned into a 50-sequence development set and a 10-sequence test set. TragicTalkers provides ground truth (GT) labels for voice activity and 2D face bounding box. The test set used for evaluation also includes 3D mouth coordinates.

3.2. Network Architecture

We tested the proposed architecture in a recent study where we tackled the sound event localization and detection (SELD) task [37]. Here, we aim to explore whether a similar network can be extended to the more specific ASDL tasks too. As depicted in Figure 3, the network presents an audio and a visual encoder to extract audio and visual feature embeddings. The embeddings

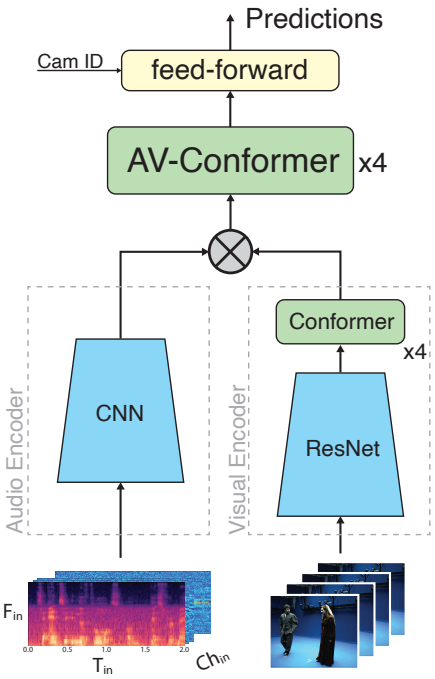

Figure 3: Proposed network architecture for audio-visual ASDL. An audio encoder based on a CNN extracts an audio embedding from the audio input features. Similarly, an encoder consisting of ResNet50 [40] followed by a Conformer unit [38] extracts a visual embedding from the video frames. $\otimes$ denotes the concatenation operation. After concatenation, the audio-visual features and processed by a second Conformer unit. A feed-forward network generates the final prediction. A camera ID one-hot vector is used to regress the speaker's position to the desired camera view.

are then concatenated and fed to an audio-visual Conformer [38]. Conformer models were originally proposed for speech recognition but lately, they have achieved state-of-the-art performance in tasks such as SELD too [39]. They present an architecture similar to the Transformer proposed by Vaswani *et al*. [31], however, they integrate a convolution module that performs pointwise and depthwise convolutions. Inspired by Wang *et al*. [39], 4 layers are employed with 8 heads each. The size of the kernel for the depthwise convolutions is set to 51, and the dimension of the hidden layer in the feed-forward networks of each Conformer layer is 1024 [39]. The output of the Conformer, i.e., the latent audio-visual representation, is then fed to a feed-forward unit consisting of three fully-connected layers to make the final predictions. A one-hot vector encoding the desired camera view of the AVA Rig is concatenated to the latent audio-visual vector after the second fully-connected layer. It allows mapping the final prediction to the desired camera of the rig. In output, the network predicts the speaker's horizontal location and voice activity confidence at a temporal resolution that matches the label and video frame rate (i.e., a position-confidence pair is generated for each video frame).

3.2.1. Audio Encoder

The audio encoder takes as input spatial features extracted from the multichannel audio signals (a detailed description of the audio in-

put features will be provided in 3.3) with shape $Ch_{in} \times T_{in} \times F_{in}$, where Ch_{in} corresponds to the number of channels of the input spatial features, T_{in} the number of temporal bins, and F_{in} the frequency bins. The audio encoder presents a CNN architecture. The CNN architecture consists of four convolutional blocks, each consisting of two 3×3 convolutional layers followed by average pooling, batch normalization [41], and ReLU activation. The average pooling layer is applied with a stride of 2, halving the temporal and frequency dimension at each block. The resulting tensor of shape $512 \times T_{in}/16 \times F_{in}/16$ is then reshaped and frequency average pooling is applied to achieve a $T_{in}/16 \times 512$ dimensional feature embedding. T_{in} is chosen so that $T_{in}/16$ matches the label frame rate of the Tragic Talkers dataset (30 labels per second).

3.2.2. Visual Encoder

As a visual encoder, the ResNet50 model [40] followed by a Conformer [38] is tested. Each video frame is fed to ResNet50. Since the video streams of Tragic Talkers are captured at 30 fps, ResNet50 extracts a number of frame embeddings that match the label frame rate as well as the audio embedding temporal resolution. The frame embeddings extracted by ResNet50 are further processed by the Conformer module, which presents the same hyper-parameters utilized for the one employed for the audio-visual fusion. The video frames used as inputs to the visual encoders are resized to 224x224p. We employed the ResNet50 model available with the torchvision library[1], which is pre-trained on image classification on the ImageNet dataset [42]. Before being fed to the Conformer, the frame embeddings are resized from the original 2048-dimensional vectors generated from ResNet50 to 512 dimensions employing a fully-connected layer to match the size of the audio embeddings.

The remainder of this paper will often refer to the visual encoder as ResNet-Conformer and to the attention-based audio-visual fusion as AV-Conformer.

3.3. Audio Input Feature

This work adopts log-mel spectrograms concatenated with generalized cross-correlation with phase transform (GCC-PHAT) features in log-mel space [43, 44] extracted from the microphone array signals. These audio features were chosen because they achieved good performance and robustness on the large microphone array of the Tragic Talker dataset [45].

The GCC-PHAT is employed to estimate the time difference of arrival (TDOA) of a sound source at two microphones [43]. The idea is to find the lag time that maximizes the cross-correlation function between the signals sensed by the two microphones. The generalized cross-correlation (GCC) is computed through the inverse Fast-Fourier Transform (inverse-FFT) of their cross-power spectrum. Phase-transformed GCC, namely the GCC-PHAT, eliminates the influence of the amplitude by leaving only the phase [44]. The GCC-PHAT between the *i*-th and the *j*-th microphone is defined at each audio frame *t* as:

$$GCC_{ij}(t,\tau) = \mathcal{F}^{-1}_{f\to\tau} \frac{\mathbf{X}_i(t,f)\mathbf{X}^*_j(t,f)}{|\mathbf{X}_i(t,f)\mathbf{X}^*_j(t,f)|}, \tag{1}$$

where $\mathbf{X}_i(t,f)$ is the Short-Time Fourier Transform (STFT) of the *i*-th channel, $\mathcal{F}^{-1}_{f\to\tau}$ the inverse-FFT from the frequency domain f to the lag-time domain τ, and $(.)^*$ denotes the complex conjugate. The TDOA can be estimated as the lag-time $\Delta\tau$ that maximizes $GCC_{ij}(t,\tau)$.

With time bins (t) on the x-axis and time-lags (τ) on the y-axis, the GCC-PHAT can be concatenated with the log-mel spectrograms extracted from the channels of the microphone array, as indicated by Cao *et al*. [44]. The concatenation of GCC-PHAT features and log-mel spectrograms provides the network with a unified representation that enables both detection and localization subtasks. We compute the GCC-PHAT between a reference microphone and the other microphones of the array. As the reference microphone, we select the first channel of the lower sub-array, as highlighted in Figure 2 (a). From the same channel, we also extract a single log-mel spectrogram for the concatenation.

3.4. Loss Function

For each input segment, the loss function $\mathcal{L}$ is determined as the sum of the individual frame losses. A sum-squared error loss [46] is computed at each output frame and is comprised of a regression and a voice activity confidence loss:

$$\mathcal{L} = \sum_{i=1}^{T_{in}/16} \mathbb{1}_i (x_i - \hat{x}_i)^2 + (C_i - \hat{C}_i)^2 \tag{2}$$

where x_i and $\hat{x}_i$ are respectively the predicted and target positions of the speaker along the horizontal axis of the i-th video frame, while C_i and $\hat{C}_i$ are the predicted and target confidences. The voice activity confidence loss is trivially achieved using the voice activity annotations: $\hat{C}_i$ is set to 1 when the frame is active and 0 when silent. The masking term $\mathbb{1}_i$ is 1 only when voice activity GT is true. It is set to 0 otherwise. So, when the frame is silent, the network is only penalized by the voice activity confidence loss and not by the regression loss. The target position of the speaker, $\hat{x}_i$, corresponds to the horizontal position of the center of the face bounding box of the active speaker, normalized by the size of the video frame to be in the range [0, 1].

4. EXPERIMENTS

4.1. Implementation Details Evaluation Metrics

The network is trained with a 5-fold cross-validation approach: each validation fold sets aside 10 unseen sequences from the 50 sequences of the development set. This cross-validation approach is used to find suitable hyper-parameters for the network. Once found, the model is retrained using the entire 50-sequence training set with these values. The network is trained for 50 epochs using batches of 32 audio feature inputs and Adam optimizer. The learning rate is fixed for the first 30 epochs, then reduced by 10% each epoch, as in [44]. The initial learning rate determined in the cross-validation is 10^{-4}.

The audio stream of the Tragic Talkers dataset is sampled at 48 kHz. The dataset is discretized into audio-visual segments of 2 seconds. The label frame rate is consistent with the video frame rate (30 fps). To align the output temporal resolution ($T_{out} = T_{in}/16$) with the labels frame rate, i.e., generating 60 activity-regression pair predictions for the 2-second input, an STFT with Hann window is applied at hop steps of 100 samples. Thus, the 2-second (96k-sample) audio chunk is discretized into 960 temporal bins (96k/100), which correspond to T_{in}. The Hann window presents

[1] https://pytorch.org/vision/stable/index.html

a size of 512 samples, as in [47]. To compute the log-mel spectrogram used in the concatenation with the GCC-PHAT features, the frequency resolution of the spectrogram is down-sampled over 64 mel-frequency bins and the logarithm operation is applied. The number of time-lags for the GCC-PHAT is also set to 64 to enable the concatenation.

The evaluation is performed on the TragicTalkers test set, labeled for 2D speaker mouth positions. A frame prediction is considered positive, i.e. the network predicts the presence of speech, when the predicted voice activity confidence is above a threshold Th, and a positive detection is true when the localization error is within a predefined spatial tolerance S. The precision and recall rates are computed by varying the confidence threshold Th from 0% to 100% sampling the thresholds from a Sigmoid-spaced distribution to provide more data points for high and low confidence values. The average precision (AP) was computed as the numerical integration of the precision-recall curve, as indicated in [48]. We set a spatial tolerance S of $\pm 2°$ along the azimuth according to human auditory perception [5], the minimum audible angle (MAA), corresponding to ± 89 pixels on the image plane. From the precision and recall rates, the F1 score is computed too. To independently evaluate the localization and the speaker detection subtasks, we define the average distance (aD) and the detection error (Det Err %) metrics. The former represents the average distance error between the active detections and the GT speaker locations. It is computed in pixels on the image frame and then converted to angle units leveraging the camera calibration data. The detection error corresponds to the percentage of frames incorrectly classified as active or inactive when a threshold $Th = 0.5$ is set.

4.2. Methods and Baselines

The proposed audio-visual method with multichannel audio is compared with two traditional audio-visual approaches for active speaker detection that employ single-channel audio and rely on face detection: Active Speaker in Context (ASC) [12] and TalkNet [11]. We also report the results achieved in our previous multichannel audio-only study [7]. The audio-only approach proposed in [7] leverages a convolutional recurrent neural network (CRNN) with an architecture similar to the audio encoder proposed in the present paper. However, instead of the Conformer unit, the CRNN presents two bidirectional gated recurrent units (biGRUs). Additionally, as a baseline system to highlight the advantages introduced by multichannel audio, we report the results achieved with a single-channel audio-only network (Mono). The input to the Mono baseline is a log mel spectrogram extracted from the central microphone of the array.

4.3. Results

The experimental results are presented in Table 1. The audio-visual multichannel (AV-M) method significantly improves the performance of the audio-only systems across nearly all metrics.

In the Mono baseline, the detection subtask is accomplished with a detection error of only 2.7%. However, the position of the speaker is not accurately predicted due to the absence of spatial cues. To minimize the error, the model locates the speaker in the central area of the frame.

Conventional active speaker detectors, such as ASC [12] and TalkNet [11], employ the audio modality only to classify the pre-extracted faces. The localisation subtask is performed by the visual face detector, yielding high spatial accuracy. However, since the average distance is achieved as the average of the predicted positions, the final average includes true and false positive predictions (false positive predictions happen when the detector classifies the silent actor as active). False positive detections are mainly present in the detections of ASC [12], penalizing its spatial accuracy as well as its detection. In the audio-visual methods with single-channel audio, the horizontal coordinate of the center of the bounding box is used as prediction, while the ground truth used for evaluation refers to the actual mouth position of the speaker. Therefore, the aD achieved is slightly overestimated due to the offset between the two representations. For example, when the speaker is captured in profile and his/her mouth is closer to the edge of the bounding box. Additionally, the AV-S methods fail when the face of the active speaker is not detected by the face detector. This causes a higher detection error and, consequently, a lower recall rate, as shown in Figure 4. At its best precision-recall pair TalkNet presents a recall rate of 79.7%, while ASC only of 58.5%. As a consequence, their overall F1 scores are affected too.

Table 1: ASDL results on the test set of the Tragic Talkers dataset and modality potential. The results for the proposed approach are achieved with audio-visual inputs and multichannel audio (AV-M). The table includes the results achieved by a single-channel audio-only network (A-S), two audio-visual systems that employ single-channel audio (AV-S), and an audio-only method that leverages multichannel audio (A-M).

Method	Mod	DetErr	aD	AP	F1
Mono	A-S	**2.7%**	210p, 4.7°	10%	30.0
ASC [12]	AV-S	43%	50p, 1.1°	59%	67.6
TalkNet[11]	AV-S	14%	35p, 0.79°	82%	84.9
CRNN [7]	A-M	3.2%	39p, 0.88°	87%	90.9
Proposed	AV-M	**2.7%**	**32p, 0.72°**	**92%**	**94.9**

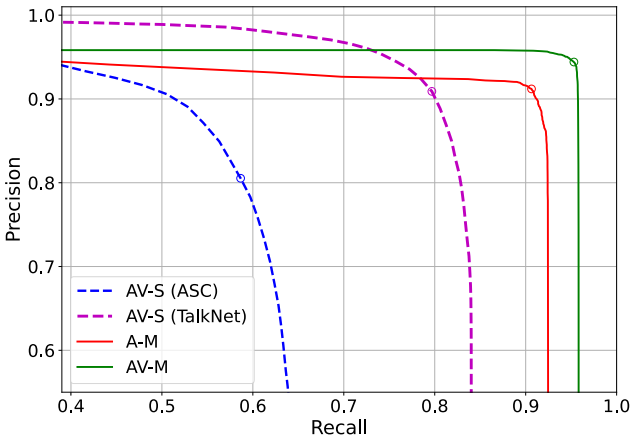

Figure 4: Comparison of precision versus recall curves for different ASDL methods. The plot includes audio-visual methods with single-channel audio (AV-S), i.e., ASC [12] and TalkNet [11], the multichannel audio-only CRNN (M-S) [7], and the proposed audio-visual approach with multichannel audio (AV-M). The combination of precision and recall rates that achieves the highest F1 score is marked on each curve.

In contrast, the multichannel audio method achieves a lower detection error as speech activity can be sensed even when the speaker is visually occluded. In fact, the double-digit detection errors of the AV-S methods are reduced to 3.2% with the A-M approach. Figure 4 shows how the gap in recall rate generated by TalkNet is halved with the multichannel audio method (90.6% recall rate). This produces an AP and F1 score higher than the AV-S systems.

When multichannel audio is partnered with vision, beneficial effects involve both detection and localization accuracy. The detection error decreases by 0.5 percentage points to 2.7%, while the aD outperforms even the audio-visual TalkNet model. The F1 score is 4 percentage points higher than the A-M approach and 10 points greater than TalkNet.

The residual error in the F1 score for the proposed AV-M method is only 5%. In the future, this error might be further narrowed by implementing visually guided predictions post-processing [49]. For example, pose detection could rectify spatial predictions using the mouth key-point coordinates. This would further improve the localization accuracy and consequently increase the number of true positive detections that fall within the spatial tolerance. Another aspect to consider is the consistency between training and testing labels. The GT labels employed to train the model correspond to center of the face bounding boxes, whereas the evaluation is based on GT mouth positions. This disparity introduces a subtle domain bias between the training and inference phases, potentially resulting in residual errors regardless of the quality of the model.

5. CONCLUSIONS

This paper proposes an audio-visual approach for active speaker detection and localization that leverages multichannel audio on the Tragic Talkers dataset. The approach extracts audio and visual embedding leveraging audio and visual encoders. Then, the embeddings are concatenated and processed by an AV-Conformer. The proposed method outperforms conventional audio-visual approaches for active speaker detection that rely on visual face detection as well as our previous audio-only multichannel work. This highlights the importance of the input modalities. Active speaker detection and localization can be employed for the automatic extraction of speaker positional metadata useful in immersive audio productions.

6. ACKNOWLEDGMENTS

This research was funded by EPSRC-BBC Prosperity Partnership 'AI4ME: Future personalised object-based media experiences delivered at scale anywhere' (EP/V038087/1). For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising. Data supporting this study are available from `https://cvssp.org/data/TragicTalkers`.

7. REFERENCES

[1] Philip Coleman, Andreas Franck, Jon Francombe, Qingju Liu, Teofilo de Campos, Richard J. Hughes, Dylan Menzies, Marcos F. Simón Gálvez, Yan Tang, James Woodcock, Philip J. B. Jackson, Frank Melchior, Chris Pike, Filippo Maria Fazi, Trevor J. Cox, and Adrian Hilton, "An audio-visual system for object-based audio: From recording to listening," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 1919–1931, 2018.

[2] Chris Pike, Richard Taylor, Tom Parnell, and Frank Melchior, "Object-based 3d audio production for virtual reality using the audio definition model," *Journal of the Audio Engineering Society*, september 2016.

[3] Mohd Azri Mohd Izhar, Marco Volino, Adrian Hilton, and Philip J. B. Jackson, "Tracking sound sources for object-based spatial audio in 3D audio-visual production," in *Forum Acusticum*, 2020, pp. 2051–2058.

[4] Florian Schweiger, Chris Pike, Tom Nixon, Matt Firth, Bruce Weir, Paul Golds, Marco Volino, Craig Cieciura, Mohd Izhar, Nick Graham-Rack, Philip J. B. Jackson, and Alex Ang, "Tools for 6-DoF immersive audio-visual content capture and production," in *International Broadcasting Convention*, 2022.

[5] Thomas Strybel and Ken Fujimoto, "Minimum audible angles in the horizontal and vertical planes: Effects of stimulus onset asynchrony and burst duration," *Journal of the Acoustical Society of America*, vol. 108 6, pp. 3092–5, 2000.

[6] Davide Berghi, Marco Volino, and Philip J. B. Jackson, "Tragic Talkers: A Shakespearean sound- and light-field dataset for audio-visual machine learning research," in *European Conference on Visual Media Production*, 2022.

[7] Davide Berghi and Philip J. B. Jackson, "Leveraging visual supervision for array-based active speaker detection and localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 984–995, 2024.

[8] Davide Berghi, Adrian Hilton, and Philip J. B. Jackson, "Visually supervised speaker detection and localization via microphone array," in *IEEE International Workshop on Multimedia Signal Processing*, 2021.

[9] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, and Caroline Pantofaru, "AVA active speaker: An audio-visual dataset for active speaker detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 4492–4496.

[10] Yuanhang Zhang, Susan Liang, Shuang Yang, Xiao Liu, Zhongqin Wu, Shiguang Shan, and Xilin Chen, "UniCon: Unified context network for robust active speaker detection," in *ACM Multimedia*, 2021.

[11] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li, "Is someone speaking? Exploring long-term temporal features for audio-visual active speaker detection," in *ACM International Conference on Multimedia*, 2021, p. 3927–3935.

[12] Juan León Alcázar, Fabian Caba, Long Mai, Federico Perazzi, Joon-Young Lee, Pablo Arbeláez, and Bernard Ghanem, "Active speakers in context," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12462–12471.

[13] Punarjay Chakravarty, Jeroen Zegers, Tinne Tuytelaars, and Hugo Van hamme, "Active speaker detection with audio-visual co-training," in *International Conference on Multimodal Interaction*, 2016, p. 312–316.

[14] Ross Cutler and Larry Davis, "Look who's talking: Speaker detection using video and audio correlation," in *IEEE International Conference on Multimedia and Expo*, 2000, pp. 1589–1592.

[15] Fasih Haider and Samer Al Moubayed, "Towards speaker detection using lips movements for human-machine multiparty dialogue," in *Fonetik*, 2012.

[16] Punarjay Chakravarty, Sayeh Mirzaei, Tinne Tuytelaars, and Hugo Van hamme, "Who's speaking? Audio-supervised classification of active speakers in video," in *International Conference on Multimodal Interaction*, 2015, p. 87–90.

[17] Ken Hoover, Sourish Chaudhuri, Caroline Pantofaru, Ian Sturdy, and Malcolm Slaney, "Using audio-visual information to understand speaker activity: Tracking active speakers on and off screen," in *IEEE International Confonference on Acoustics, Speech and Signal Processing*, 2018, pp. 6558–6562.

[18] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman, "VoxCeleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, pp. 101027, 2020.

[19] Joon Son Chung and Andrew Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*, 2016.

[20] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "ASR is all you need: Cross-modal distillation for lip reading," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 2143–2147.

[21] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang, "Perfect match: Improved cross-modal embeddings for audio-visual synchronisation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 3965–3969.

[22] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "The conversation: Deep audio-visual speech enhancement," in *INTERSPEECH*, 2018.

[23] Joon Son Chung, "Naver at ActivityNet Challenge 2019 - Task B Active speaker detection (AVA)," *ArXiv*, vol. abs/1906.10555, 2019.

[24] Yuan-Hang Zhang, Jing-Yun Xiao, Shuang Yang, and Shiguang Shan, "Multi-task learning for audio-visual active speaker detection," in *Techical Report AVA-ActiveSpeaker Challenge*, 2019.

[25] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," *ArXiv*, vol. abs/1809.00496, 2018.

[26] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[27] Juan León Alcázar, Moritz Cordes, Chen Zhao, and Bernard Ghanem, "End-to-end active speaker detection," in *European Conference on Computer Vision*, 2022, p. 126–143.

[28] Xinyuan Qian, Maulik Madhavi, Zexu Pan, Jiadong Wang, and Haizhou Li, "Multi-target DoA estimation with an audio-visual fusion mechanism," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 4280–4284.

[29] Xinyuan Qian, Zhengdong Wang, Jiadong Wang, Guohui Guan, and Haizhou Li, "Audio-visual cross-attention network for robotic speaker tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 550–562, 2023.

[30] Yulin Wu, Ruimin Hu, Xiaochen Wang, and Shanfa Ke, "Multi-speaker DoA estimation using audio and visual modality," *Neural Processing Letters*, vol. 55, no. 7, pp. 8887–8901, 2023.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *International Conference on Neural Information Processing Systems*, 2017, p. 6000–6010.

[32] Jinzheng Zhao, Peipei Wu, Shidrokh Goudarzi, Xubo Liu, Jianyuan Sun, Yong Xu, and Wenwu Wang, "Visually assisted self-supervised audio speaker localization and tracking," in *European Signal Processing Conference*, 2022, pp. 787–791.

[33] Jinzheng Zhao, Yong Xu, Xinyuan Qian, and Wenwu Wang, "Audio visual speaker localization from egocentric views," *ArXiv*, vol. abs/2309.16308, 2023.

[34] Jacob Donley, Vladimir Tourbabin, Jung-Suk Lee, Mark Broyles, Hao Jiang, Jie Shen, Maja Pantic, Vamsi Krishna Ithapu, and Ravish Mehra, "Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments," *arXiv*, vol. 2107.04174, 2021.

[35] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu, "Egocentric deep multi-channel audio-visual active speaker localization," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10534–10542, 2022.

[36] Ilya Gurvich, Ido Leichter, Dharmendar Reddy Palle, Yossi Asher, Alon Vinnikov, Igor Abramovski, Vishak Gopal, Ross Cutler, and Eyal Krupka, "A real-time active speaker detection system integrating an audio-visual signal with a spatial querying mechanism," *ArXiv*, vol. abs/2309.08295, 2023.

[37] Davide Berghi, Peipei Wu, Jinzheng Zhao, Wenwu Wang, and Philip J. B. Jackson, "Fusion of audio and visual embeddings for sound event localization and detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.

[38] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented transformer for speech recognition," *ArXiv*, vol. abs/2005.08100, 2020.

[39] Qing Wang, Jun Du, Hua-Xin Wu, Jia Pan, Feng Ma, and Chin-Hui Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1251–1264, 2023.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[41] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, vol. 37, pp. 448–456.

[42] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li, "ImageNet: A large-scale hierarchical image database," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[43] Charles Knapp and G. Clifford 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.

[44] Yin Cao, Qiuqiang Kong, Turab Iqbal, Fengyan An, Wenwu Wang, and Mark D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Detection and Classification of Acoustic Scenes and Events Workshop*, 2019.

[45] Davide Berghi and Philip J. B. Jackson, "Audio inputs for active speaker detection and localization via microphone array," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2023.

[46] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[47] Thi Ngoc Tho Nguyen, Karn N. Watcharasupat, Ngoc Khanh Nguyen, Douglas L. Jones, and Woon-Seng Gan, "SALSA: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 1749–1762, 2022.

[48] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman, "The PASCAL visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.

[49] Qing Wang, Ya Jiang, Shi Cheng, Maocheng Hu, Zhaoxu Nian, Pengfei Hu, Zeyan Liu, Yuxuan Dong, Mingqi Cai, Jun Du, and Chin-Hui Lee, "The NERC-SLIP system for sound event localization and detection of DCASE2023 challenge," in *Techical Report of DCASE Challenge*, 2023.