
Proceedings Paper:
Clarke, J., Gotoh, Y. and Goetze, S. (2024) Improving audiovisual active speaker detection in egocentric recordings with the data-efficient image transformer. In: 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2023),, 16-20 Dec 2023, Taipei, Taiwan. Institute of Electrical and Electronics Engineers (IEEE). ISBN: 979-8-3503-0690-3.

https://doi.org/10.1109/ASRU57964.2023.10389764

IMPROVING AUDIOVISUAL ACTIVE SPEAKER DETECTION IN EGOCENTRIC RECORDINGS WITH THE DATA-EFFICIENT IMAGE TRANSFORMER

Jason Clarke, Yoshihiko Gotoh, and Stefan Goetze

Department of Computer Science, The University of Sheffield, Sheffield, United Kingdom

ABSTRACT

Future augmented reality devices have the capacity to enhance human perception and provide assistive functions in complex communication scenarios. Active speaker detection (ASD) systems that are robust to egocentric data are critical to this. Egocentric ASD is challenging due to overlapping speech, single-channel recording, and dynamic scenes. A novel module that uses a data-efficient image transformer (DeiT) to extract features encapsulating the acoustic properties of each scene, and a positional conditioning mechanism is proposed. The module is evaluated in conjunction with TalkNet, an existing ASD architecture, on two audiovisual datasets: Ego4D (egocentric) and AVA-ActiveSpeaker (exocentric), achieving 29% and 0.38% relative improvement in mean Average Precision (mAP), respectively, while retaining a parameter efficient build. A qualitative analysis is also presented, implicitly demonstrating that contextual information is leveraged.

***Index Terms*—** Active speaker detection, context modelling, data-efficient image transformers

1. INTRODUCTION

The ability to determine the identity of an active speaker in a communication scenario is a fundamental aspect of human social interaction. Audiovisual diarization aims to identify and associate speech segments with the relevant identities present in an audiovisual signal. Audiovisual ASD is a critical component of modular audiovisual diarization frameworks. The objective of these frameworks is to detect all active speakers present in an audiovisual scene given contiguous bounding boxes of faces and the corresponding audio. This binary classification is often performed at a video-frame level of temporal granularity [1–4].

Future applications of ASD in the context of augmented reality revolve around data acquired from the egocentric perspective. Egocentric refers to video or audio recorded from the first-person perspective, typically by wearable devices such as smart glasses. Despite this, most existing ASD methods are designed for, and evaluated using, benchmarks consisting of conventional exocentric audio and video captured by microphones and static cameras in relatively favourable recording environments [4, 5]. The Ego4D [6] egocentric audiovisual dataset and benchmark suite has provided a suitable framework to study ASD for egocentric data in detail. Such data introduces a notably arduous set of challenges: (i) acoustic signal level differences between the signal of the camera wearer and their interlocutors due to distance differences to the microphone; (ii) generally low signal-to-noise-ratio (SNR) for speech signals; (iii) significant visual distortion from motion blurring induced by the dynamics of the camera wearer's head movement; (iv) adverse lighting conditions; and (v) prevalent occurrences of overlapping speech due to highly spontaneous conversations.

For ASD systems which scrutinise the activity of a single candidate speaker in isolation, these challenges prove particularly difficult [6, 7]. To mitigate similar difficulties present in exocentric datasets, such as the standard ASD benchmark AVA-ActiveSpeaker [4], recent research in the domain has opted for approaches which leverage information from the context, e.g. provided by the wider image of each video-frame [1, 2, 8–11]. Specifically, most work has focused on leveraging information provided by the facial regions of visible interlocutors surrounding the candidate speaker. This is typically done in two ways: either by determining the activity of all candidate speakers within a frame simultaneously [2, 9, 10], or simply by using them as a source of contextual information [1]. Other literature has taken a step further by incorporating the position [8] and physical size of each speaker's head [11] within the image. This is based on the heuristic that the active speaker is more likely to be located in the center of the image and having a larger head size than inactive speakers.

Beyond these features, an intuitively relevant piece of contextual information has previously been overlooked. By modelling each video-frame holistically, the specific environment in which the scene encapsulates can act as a relevant prior; the acoustic properties of the environment can be inferred from the whole image of each video frame. This is in addition to other features which can only be inferred from viewing extended portions of the whole image (beyond conventional facial crops of speakers), such as the body language and orientation of the visible interlocutors surrounding the candidate speaker.

This work presents a novel method to leverage the contextual information provided by the full-scene image of each video frame using a pretrained DeiT [12] and a positional conditioning mechanism. TalkNet, an existing ASD architecture which detects the

THIS WORK WAS SUPPORTED BY THE CENTRE FOR DOCTORAL TRAINING IN SPEECH AND LANGUAGE TECHNOLOGIES (SLT) AND THEIR APPLICATIONS FUNDED BY UKRI [GRANT NUMBER EP/S023062/1]. THIS WORK WAS ALSO FUNDED IN PART BY META.

speech activity for a single candidate speaker in isolation (and is therefore naive to any contextual information) is used as a baseline to test the efficacy of said extension. This work finds that the proposed extension yields significant improvement upon the TalkNet baseline on the egocentric dataset Ego4D as well as a modest improvement on the AVA-ActiveSpeaker exocentric benchmark.

Contributions:

1. A novel context modelling extension for audiovisual ASD to disambiguate acoustically challenging scenes in egocentric data, all trained models and code are publicly available[1].
2. Experiments on both Ego4D and AVA-ActiveSpeaker datasets which demonstrate the efficacy of the proposed extension: 29% and 0.38% relative improvement compared to the TalkNet baseline are acheived, respectively.
3. Qualitative analysis of the performance, including an ablation study and simulations to demonstrate how the proposed extension manages populated audiovisual scenes.

The remainder of this work is organised as follows: Section 2 provides the necessary theoretical background and motivates the architectures used for this work. Section 3 describes the proposed method to model the context provided by each full-scene image. Section 4 provides an overview of datasets used and implementation details. Section 5 discusses experimental results, first a comparison with the state-of-the-art in both ego- and exocentric data, then a qualitative analysis, and Section 6 concludes the paper.

2. RELATED WORK

The two architectures fundamental to this work will be introduced in the following, i.e. the audiovisual ASD system with relevant notation and a technical description of the TalkNet audiovisual ASD baseline architecture. Then, the DeiT architecture [12] is briefly introduced and motivated as a context modelling mechanism to assist in disambiguating acoustically challenging environments.

2.1. Notation and Overview

Fig. 1 shows the audiovisual ASD system to determine the video-frame-wise speech activity of a candidate speaker S who is visibly present uninterrupted in a set $\mathcal{V} = \{\mathbf{V}^1, ..., \mathbf{V}^T\}$ of consecutive video frames $\mathbf{V}^t \in \mathbb{R}^{C \times H \times W}$ with time index $t \in \{1, ..., T\}$ and C, H, and W being the channels, height, and width dimensions of each full-scene image, respectively. First, contiguous bounding boxes encapsulating the facial region of the candidate speaker $\mathcal{V}_\mathrm{S} = \{\mathbf{V}_\mathrm{S}^1, ..., \mathbf{V}_\mathrm{S}^T\}$ are extracted from each full-scene image. These contiguous bounding boxes are typically referred to as streams or tracklets. A temporally corresponding audio signal $\mathcal{A} = \{\mathbf{a}^1, ..., \mathbf{a}^{T_\mathrm{A}}\}$ is used in conjunction with V_S to infer the speech activity of the candidate speaker S. It is worth noting that $\mathcal{A}$ has a time index $t_\mathrm{A} \in \{1, ..., T_\mathrm{A}\}$ which is distinct from t, to account for the discrepancy in modality sampling rates.

[1] https://github.com/sap-shef/full_scene_ASD

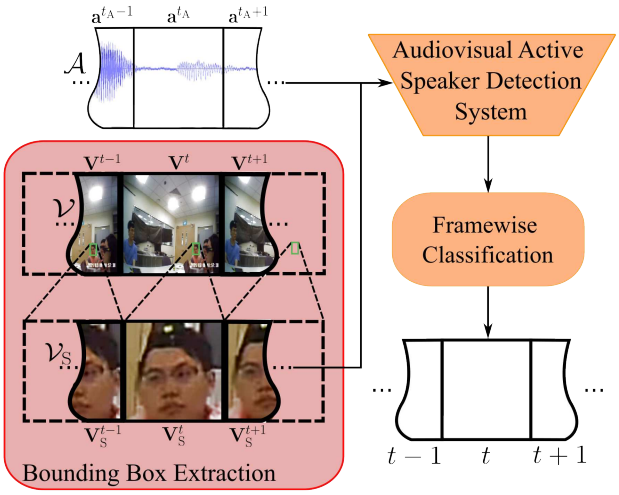

Fig. 1. Overall framework of an audiovisual ASD system.

Unsupervised and supervised approaches exist for ASD. For example, one such unsupervised approach assumes the active speaker for segment $\mathcal{A}$ is the speaker whose $\mathcal{V}_\mathrm{S}$ and $\mathcal{A}$ have the highest audio-visual alignment [13, 14]. This is based on the assumption that speech present in the audio signal *must* belong to one of the speakers visibly present in the field of view (FoV). Another approach is based on the premise that the faces that co-occur most frequently with pre-diarized speaker identities can be matched together [15]. This is based on the assumption that speaker identities can be robustly diarized beforehand; it therefore cannot be applied causally. Comparatively, supervised approaches formulate the task as a binary classification problem [1, 2, 4, 7, 10, 11], with fewer assumptions. Adversities encountered when dealing with *in the wild* data often make the aforementioned assumptions invalid. Thus, recent literature has observed the paradigm shift to supervised approaches.

2.2. TalkNet Baseline

This paper proposes a novel approach to enhance the performance of ASD in the context of acoustically challenging scenes in egocentric video by injecting contextual information from each full-scene image. To evaluate the proposed approach, TalkNet [7] is chosen as an existing baseline system that performs ASD without any contextual information.

TalkNet follows the paradigm of comprising an audio encoder, a video encoder, modality fusion, and a temporal modelling mechanism [1,2,4,8–11,16]. The audio encoder embeds 13-dimensional Mel frequency cepstral coefficients (MFCCs) using a ResNet-34 with squeeze and excitation layers [17], in conjunction with a static receptive field and a dynamic MFCC window step [7]. The dynamic MFCC window step accounts for the discrepancy in modality sampling rates and dynamic input sizes. The video encoder uses a ResNet-18 to model the spatial features of the candidate speaker followed by a video temporal convolution module that consists

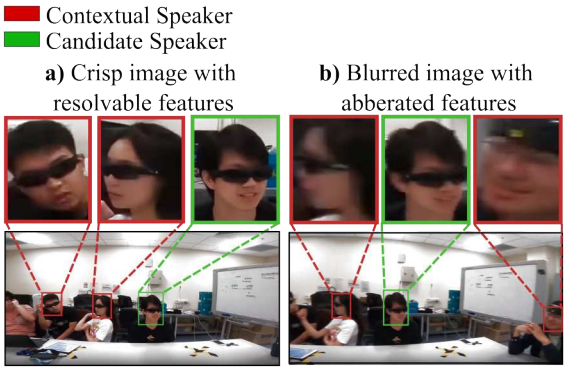

Fig. 2. Left: Crisp image; visual features are indicative of speaker activity, e.g. open mouths resolvable. Right: Blurred image where it is difficult to resolve such features. Blurring occurs across both candidate speaker, and contextual speakers (surrounding interlocutors). Images from Ego4D dataset [6].

of depth-wise separable convolutions across the temporal dimension [18]. For modality fusion, two cross-attention mechanisms perform audiovisual synchronization by aligning the two modalities followed by embedding-dimension concatenation. To model the interframe dynamics of an encoded input stream TalkNet makes use of the self-attention mechanism. This enables TalkNet to infer from the full stream for temporal context when classifying the activity of each audio and video frame as well as permitting dynamic sequence lengths of input. Maximising the windows of temporal context has been shown to be beneficial to ASD performance [7, 8] and the duration of a stream is often variable.

As a result of its strong modality synchronisation capabilities and the long windows of temporal context it can infer from, TalkNet retains competitive performance in exocentric ASD despite being several years older and assuming a parameter-efficient approach relative to more recent work [1, 10, 11, 16]. Additionally, TalkNet does not require significant architectural modification to incorporate the proposed visual context modelling extension. This ensures that its original mechanism for the candidate speaker-specific visual features and audio is not significantly disrupted by the proposed extension. These factors make it the optimal choice as a baseline framework.

2.3. Data Efficient Image Transformers

Several challenges exist for ASD systems in the context of egocentric video. From an audio perspective, noisy environments and frequent occurrences of overlapping speech make it difficult to determine whether speech present in the audio signal emanates from the candidate speaker in question, or, whether speech is present in the audio signal at all [19, 20]. Additionally, the visual degradation which inevitably occurs due to the dynamics of the camera wearer's head movement poses a challenge from a visual perspective. This motion often results in significant distortion to the candidate speakers' face crop, rendering the fine-grained details typically associated with speech activity as unreliable; it is difficult to recognise an open mouth or cheek posture [21] under such distortion. To mitigate these challenges, other contextual information provided by the full-scene image can be leveraged. Using the full-scene image, a system can be informed of potential overlapping speech and noisy environments by identifying other visible persons and inspecting the scene background, respectively (cf. Fig. 2). Furthermore, inferring from a full-scene image holistically is advantageous because the scene-level information is less susceptible to visual distortion. For example in Fig. 2, elementary features such as the interlocutor position, body language, and the environment the scene encapsulates can still be discerned even under the presence of aberration. To leverage robust scene-level information, this work proposes the use of a DeiT [12] as a means of modeling the visual context provided by each full-scene image to assist audiovisual ASD.

The DeiT is a type of vision transformer [22] that leverages self-attention to process images. Unlike convolutional neural networks, vision transformers can capture long-range dependencies and global features in images, with weaker inductive biases [23]. However, vision transformers require a large amount of data and computing resources to train, which limits their adoption in lightweight ASD architectures. The DeiT addresses this issue by using a teacher-student strategy that relies on a distillation token, which ensures that the student learns from the teacher through attention. Pretrained on the ImageNet-1K dataset [24], the DeiT can be finetuned with modest resources for a wide range of tasks. This paper demonstrates that the DeiT can autonomously extract contextual information from full-scene images for ASD.

3. CONTEXT MODELLING

This section describes the proposed method to extend TalkNet [7], which does not consider visual context, with a visual context modelling module. The proposed extension aims to improve ASD performance by injecting contextual information to compensate for audible and visual noise inherent to egocentric data. The upper part in Fig. 3 shows the extension and the lower part shows the baseline architecture, the latter is identical to its original implementation (as described in Section 2.2).

The baseline architecture first encodes a stream of the local visual features of the candidate speaker $\mathcal{V}_\mathrm{S}$ and the audio signal $\mathcal{A}$ yielding video and audio representations $\mathbf{F}_\mathrm{V} \in \mathbb{R}^{T\times d}$ and $\mathbf{F}_\mathrm{A} \in \mathbb{R}^{T\times d}$, respectively. Here d denotes the embedding dimension. The embedded modalities are then concatenated to $\mathbf{F}_\mathrm{AV} \in \mathbb{R}^{T\times 2d}$. The extension extracts information from the corresponding stream of full-scene images $\mathcal{V}$ and conditions the embedded representations by the position of the candidate speaker within each full-scene image $\mathcal{P}_\mathrm{S} = \{\mathbf{p}_\mathrm{S}^1,...,\mathbf{p}_\mathrm{S}^T\}$, given $\mathbf{p}_\mathrm{S}^t = [x_1,y_1,x_2,y_2] \in \mathbb{R}^{4\times 1}$ where x_1, y_1, x_2, and y_2 denote the upper left and lower right coordinates of the bounding box (shown as orange dots in Fig. 3). This positionally conditioned representation of each full-scene image $\mathbf{F}_\mathrm{C} \in \mathbb{R}^{T\times d_\mathrm{c}}$ is then concatenated with $\mathbf{F}_\mathrm{AV}$ along the embedding dimension, yielding an embedded representation of the full

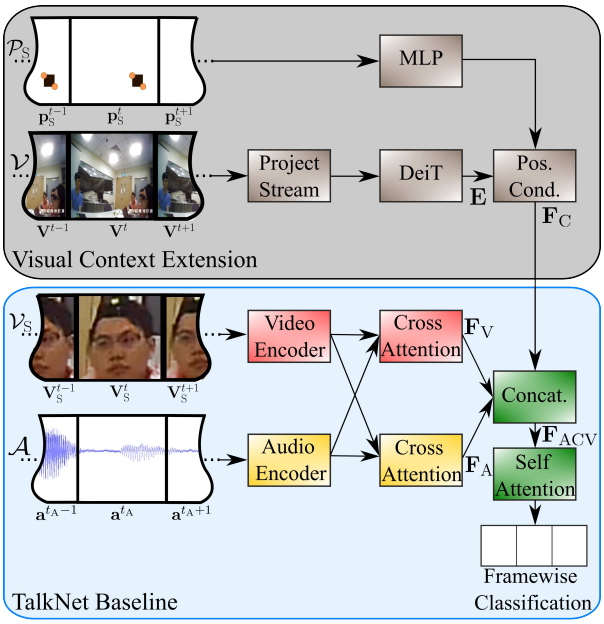

Fig. 3. Visual context modelling extension (upper part) and TalkNet [7] baseline (lower part) for ASD.

stream $\mathbf{F}_{\mathrm{ACV}} \in \mathbb{R}^{T\times(2d+d_{\mathrm{c}})}$ from each input. As per the baseline architecture, to infer from the full temporal context, self-attention is then applied across the entire embedded stream. Finally, the network classifies each video frame within the embedded stream as either being representative of an active speaker or not.

3.1. Full-Scene Feature Extraction

Most ASD systems use contextual information only from the faces of other visible interlocutors surrounding the candidate speaker [1, 2, 9]. This can work well in some situations, like when the audio signal emanates from *one* of the non-distorted faces in the FoV [2]. However, this approach is not robust to the challenges of egocentric video. For example, if the candidate speaker's face is blurred by camera motion, the faces of the surrounding interlocutors are also likely to be blurred. This makes it difficult to leverage useful contextual information that indicates the identity of the active speaker. Therefore, in the context of egocentric data, this kind of contextual information is susceptible to the same affliction it is trying to mitigate, as demonstrated in Fig. 2. Moreover, this approach restricts the amount of information the model can extract from the visual context regarding the audio signal. It is limited to being informed of the possibility of overlapping speech. If multiple interlocutors are visible, this would increase the likelihood that the audio signal contains overlapping speech. A more informed approach would exploit information regarding the specific environment each full-scene image encapsulates and identify potential sources of audible sound. To this end, this work proposes to model each full-scene image holistically using a DeiT.

First, each full-scene image within a stream is stratified into a mosaic of fixed-size patches where a single patch is $\mathbf{\Phi} \in \mathbb{R}^{C\times\frac{H}{P_H}\times\frac{W}{P_W}}$ given P_H and P_W denoted the total number of patches in each image in height and width direction, respectively. Each patch is spatially flattened and projected into a vector $\mathbf{\Phi}_{\mathrm{emb}} \in \mathbb{R}^{d_{\mathrm{c}}}$ (where d_{c} refers to the model dimension of the DeiT) via linear layers, resulting in a matrix $\mathbf{V}^t_{\mathrm{emb}} \in \mathbb{R}^{P\times d_{\mathrm{c}}}$ for each full-scene image in the stream. The matrices of all the images in the stream are stacked along the temporal dimension, forming the tensor $\mathcal{V}_{\mathrm{emb}} \in \mathbb{R}^{T\times P\times d_{\mathrm{c}}}$. Subsequent transformer blocks comprising self-attention and feed-forward layers with residual connections are then used to autonomously extract information about each image within the stream, yielding a tensor $\mathbf{E} \in \mathbb{R}^{T\times P\times d_{\mathrm{c}}}$.

3.2. Positional Conditioning

Previous work demonstrates that elementary contextual information such as the position of the candidate speaker within each full-scene image is insightful when detecting active speakers [8, 11]. To further build on this idea, the use of a positional conditioning mechanism is proposed. The goal of this mechanism is to further enrich the full-scene image embedding obtained by the DeiT, by conditioning it on the position of the candidate speaker using a cross-attention mechanism. This is based on the premise that regions of salience within the full-scene image, as determined by the DeiT, can be further weighted in terms of their relevance by the position of the candidate speaker. For example, patches within the image containing persons will be identified by the DeiT as salient. However, to leverage information within those patches, such as the head orientation of the recognised person, the position of the candidate speaker must be known. Additionally, the distance of potential sources of audible noise from the camera wearer relative to the distance to the candidate speaker is also a useful prior. Relationships such as these can be more easily learnt using a cross-attention mechanism compared to other, simpler methods.

First, a multilayer perceptron (MLP) is used to embed the positions of the candidate speaker within each full-scene image among the stream $\mathbf{E}_{\mathrm{p}} \in \mathbb{R}^{T\times 1\times d_{\mathrm{D}}}$. Cross-attention is then applied, as per (1), where the sequence length of the keys and values is the patch dimension of $\mathbf{E}$, and the sequence length of the queries is simply one.

$$\mathrm{ATT}(\mathbf{E}) = \mathrm{softmax}\left(\frac{\mathbf{E}_{\mathrm{p}}\mathbf{E}^{\intercal}}{\sqrt{d_{\mathrm{D}}}}\right)\mathbf{E} \tag{1}$$

The cross-attention mechanism weights each patch by its relevance to the task of determining the candidate speaker's activity for the corresponding video frame. To obtain the final context embedding, the patch dimension is collapsed by mean average, yielding $\mathbf{F}_{\mathrm{C}} \in \mathbb{R}^{T\times d_{\mathrm{D}}}$.

4. EXPERIMENTS

Ego4D. The model is first pretrained on the train fold of the AVA-ActiveSpeaker dataset [4]. Next, 5 separate runs are trained for 25 epochs on the train fold of the Ego4D audiovisual diarization

dataset. For augmentation, during training, negative sampling [7] is applied to the audio signal, and standard techniques such as random flipping, rotating, and cropping are applied to the video modality. For evaluation, the Cartucho implementation of object detection mAP [25] is used. This follows the mAP criterion defined in the PASCAL VOC 2012 competition [26], and is the same evaluation protocol provided by the Ego4D audiovisual diarization challenge which ensures evaluatory consistency with recent literature [6].

AVA-ActiveSpeaker. The model is trained on the AVA-ActiveSpeaker dataset also for 25 epochs with 5 distinct runs using the augmentation protocol described above. The optimal checkpoint is then evaluated using the official evaluatory code provided by the ActivityNet challenge [27].

4.1. Implementation Details

The PyTorch implementation of the DeiT-tiny-patch16-224 from the Facebook repository on Hugging Face is used as the pretrained DeiT. It takes 224×224 colour images as input. The TalkNet baseline architecture is consistent with its original implementation [7], it uses 13-dimensional MFCCs as input to the audio encoder with a receptive field of 189 audio frames with dynamic MFCC window steps. The video encoder accepts 112×112 images of the candidate speaker's face crop in grey-scale. The embedding dimensions for $\mathbf{F}_{\mathrm{A}}$, $\mathbf{F}_{\mathrm{V}}$, and $\mathbf{F}_{\mathrm{C}}$ are 128, 128, and 64, respectively. The loss function used is the weighted cross entropy loss to compensate for the class bias of the data set, with 3 auxiliary losses: audio, visual, and contextual. It was determined empirically that a larger contextual loss weighting relative to the audio and visual loss contributions yields better results, therefore, the auxiliary losses were weighted as follows: 0.4, 0.4, and 0.7. The temporal context the model can infer from is dynamic, ranging from 2 to 300 video-frames (0.67 to 10 seconds assuming a 30 Hz video sampling rate). Since the test folds of Ego4D and AVA-ActiveSpeaker are unavailable, i.e. have not been released by the Ego4D audiovisual diarization challenge and ActivityNet challenge, respectively, this work follows other literature [1, 7, 8, 10, 16] and uses the validation folds of each dataset for evaluation. The full model is trained in less than 24 hours using a single V100 and a batch size of 900 video frames.

4.2. Datasets

4.2.1. Ego4D Dataset

Ego4D [6] is a large dataset of over 3000 hours of annotated video recorded from the egocentric perspective. It covers tasks from episodic memory to audiovisual diarization. This work, however, only uses the audiovisual diarization component of the dataset which comprises 572 distinct video clips. Each video clip is 5 minutes long, some of which are recorded concurrently. All data is recorded monaurally using a variety of wearable devices. All video is sampled at 30 Hz and uses high-definition visual resolution. The dataset is stratified as follows: 379 clips for training, 50 clips for validation, and 133 clips for testing. The dataset records real-life conversational scenarios, usually involving multiple speakers, both indoor and outdoor. The dataset is therefore incredibly diverse and particularly challenging.

4.2.2. AVA-ActiveSpeaker Dataset

The AVA-ActiveSpeaker dataset [4] is the first large-scale standard benchmark for ASD, with 262 exocentrically recorded Hollywood movie clips as its source. The dataset is split as follows: 120 training movie clips, 33 validation movie clips, and 109 test movie clips. Faces are annotated to a video-frame level of temporal granularity for speaker activity, yielding bounding boxes for 5.3 million faces. The dataset poses various challenges for ASD, such as occlusions, low-resolution faces, low-quality audio, and challenging lighting conditions.

5. RESULTS

5.1. Comparison with State-of-the-Art Methods

Primarily the objective of this work is to enhance the performance of audiovisual ASD in the context of egocentric data. However, for completeness, it is necessary to assess the performance of the system in the context of exocentric data as well. To this end the proposed system is evaluated on the two datasets described in Section 4.2: the Ego4D audiovisual diarization dataset and the AVA-ActiveSpeaker dataset.

Table 1. Performance on validation set of the AVA-ActiveSpeaker dataset. Symbol $>$ refers to a minimum estimate of parameters present in the model. *Cont. Inf.* denotes whether the model is contextually informed. Best performance in bold font.

Model	Cont. Inf.	mAP [%]	Params. [M]
AVA [4]	✗	82.1	>10.0
Zhang et al.	✗	83.5	>35.0
ASC [2]	✓	87.1	23.5
MAAS [9]	✓	88.1	22.5
UniCon [11]	✓	92.2	>22.4
TalkNet [7]	✗	92.3	15.7
Proposed Method	✓	92.7	21.2
ASDNet [1]	✓	93.5	51.3
Liao et al. [28]	✗	94.1	**1.00**
SPELL [8]	✓	94.2	22.5
LoCoNet [16]	✓	**95.2**	>22.5

When evaluated on the AVA-ActiveSpeaker dataset, as shown in Table 1, the improvement upon the TalkNet baseline is modest, resulting in a 0.38% relative improvement in mAP. This might be for two reasons. Firstly, the challenges the proposed extension helps to mitigate such as overlapping speech, acoustic noise, and visual distortion are less prevalent in exocentric data. Secondly, the baseline performance is already comparable with the current state-of-the-art, leaving less room for improvement. Furthermore, due to the subjective nature of the annotation, some disagreement

with the ground-truth ASD labels is likely to occur even with a perfect ASD system. Therefore, diminishing returns are expected beyond a certain threshold of improvement.

Table 2. Performance on Ego4D-val [6]. Results reported are taken from existing literature, except for Liao et al [28]. For LoCoNet, a minimum number of parameters is stated since the number of parameters used is variable. All systems shown, excluding the TalkNet baseline and Liao et al., are contextually informed.

Model	Params. [M]	mAP [%]
TalkNet [7]	15.7	51.0
Liao et al. [8]	**1.00**	54.3
LoCoNet [16]	>22.5	59.7
SPELL [8]	22.5	60.7
Proposed Method	21.2	**65.9**

In the context of egocentric data, as shown in Table 2, the proposed extension significantly improves upon the TalkNet baseline system, yielding a 29% relative improvement in mAP. Additionally, it significantly outperforms the current state-of-the-art system on the Ego4D dataset, SPELL [8], by 5.2% mAP. This is whilst using less parameters than SPELL which, unlike the proposed method, is also not trained in an end-to-end fashion and is non-causal. These results implicitly validate the hypothesis of this work; contextual information provided by the full-scene image is beneficial when determining speaker activity. This is particularly true in the case of egocentric data.

5.2. Ablation Study

To implicitly determine the kinds of information the context modelling extension learns from, an ablation study is conducted in the following. In total four adaptations of the context modelling extension are evaluated with the TalkNet baseline: configuration (i) full-scene image† + position refers to the proposed extension, but using full-scene images where all regions except for those including face crops have been ablated and replaced with pixel values of zero; configuration (ii) position-only refers to positional information being injected without any full-scene image embeddings; configuration (iii) full-scene image, where the positional conditioning mechanism has been ablated; and finally, configuration (iv) full-scene image + position which is the proposed extension to TalkNet, unmodified. From Table 3 it is clear that to effectively leverage the contextual information provided by the full-scene image, it is necessary to extract holistic information and condition the embedded representation on the position of the candidate speaker. Surprisingly, configuration (i) performs the worst. This may be due to the fact that the majority of the embedded patches in this configuration simply contain no information, since visible faces only consume a small minority of area within each full-scene image. This means the model is inundated with noise and has to learn how to selectively ignore certain patches. This experiment also confirms that using the full-scene image to capture the acoustic features of the environment can help identify the active speakers in egocentric data.

Table 3. Ablation study for visual context modelling components evaluated on Ego4D-val [6]. † indicates ablated full-scene images only comprising the bounding box regions of visible interlocutors.

Model	mAP [%]
full-scene image† + position	60.4
position-only	60.5
full-scene image	61.3
full-scene image + position	**65.9**

5.3. Qualitative Analysis

Table 4 shows a breakdown of the baseline performance compared with the proposed method in terms of mAP, stratified by the number of visible interlocutors in each scene. For all strata, excluding scenes comprising 3 visible interlocutors, the proposed method produces a significant improvement. In previous studies [1, 2, 4, 7, 9, 28], a clear trend of decreasing performance with increasing number of visible interlocutors in each scene is apparent. This trend is consistent with what the baseline system exhibits on the Ego4D validaiton fold. The proposed method, however, does not abide by this trend, performance increases with increasing visible interlocutors beyond 3. This may be because information provided by visible interlocutors in a scene, such as head orientation, is beneficial when determining speaker activity. The proposed extension potentially leverages this information allowing it to perform better in particularly crowded scenes. For this phenomena to overcome the challenges induced by crowded scenes, there must be a significant quantity of visible interlocutors; it is unlikely that all interlocutors are looking at the active speaker at all times. For example, in a scene comprising 3 visible interlocutors, there is an insufficient number of visible interlocutors for their gaze to not provide an ambiguous result. This would explain the dip in performance for scenes comprising 3 visible interlocutors, observed in Table 4.

Table 4. Effect of number of visible interlocutors in the FoV on the detection performance in terms of mAP and scenario abundance.

# of Visible interlocutors	1	2	3	4	≥5
TalkNet	58.1	39.4	**22.3**	22.6	16.0
Proposed Method	**71.9**	**58.0**	20.3	**25.1**	**47.5**
Abundance [% of val fold]	46	52	0.70	0.63	0.0072

6. CONCLUSIONS

In this study, a context modelling module to extend an existing audiovisual ASD system to improve its performance in particular for egocentric data was proposed. The extension uses a pretrained DeiT and a positional conditioning mechanism to extract and leverage contextual information reflecting the acoustic properties of full-scene images. Experimentation on two ASD benchmarks demonstrate that the proposed extension achieves 65.9% and 92.7% on egocentric and exocentric data, respectively, significantly outperforming all other methods on the Ego4D dataset.

7. REFERENCES

[1] O. Köpüklü, M. Taseska, and G. Rigoll, "How to design a three-stage architecture for audio-visual active speaker detection in the wild," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1173–1183.

[2] J. L. Alcazar, F. C. Heilbron, L. Mai, F. Perazzi, J.-Y. Lee, P. Arbeláez, and B. Ghanem, "Active speakers in context," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12462–12471, 2020.

[3] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with lstms for lipreading," *ArXiv*, vol. abs/1703.04105, 2017.

[4] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi, and C. Pantofaru, "Ava Active Speaker: An Audio-Visual Dataset for Active Speaker Detection," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[5] P. Chakravarty and T. Tuytelaars, "Cross-modal supervision for learning active speaker detection in video," in *European Conference on Computer Vision*, 2016.

[6] "Ego4d: Around the world in 3,000 hours of egocentric video," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18973–18990, 2021.

[7] 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," in *Proc. 29th ACM Int. Conf. on Multimedia*, 2021, p. 3927–3935.

[8] K. Min, S. Roy, S. Tripathi, T. Guha, and S. Majumdar, "Learning long-term spatial-temporal graphs for active speaker detection," in *Euro. Conf. on Computer Vision*, 2022.

[9] J. Le'on-Alc'azar, F. C. Heilbron, A. K. Thabet, and B. Ghanem, "Maas: Multi-modal assignation for active speaker detection," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 265–274, 2021.

[10] J. L. Alcazar, M. Cordes, C. Zhao, and B. Ghanem, "End-to-end active speaker detection," in *European Conference on Computer Vision*, 2022.

[11] Y. Zhang, S. Liang, S. Yang, X. Liu, Z. Wu, S. Shan, and X. Chen, "Unicon: Unified context network for robust active speaker detection," *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[12] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *Proc. 38th Int. Conf. on Machine Learning*, M. Meila and T. Zhang, Eds., 2021, vol. 139 of *Proc. Machine Learning Research*.

[13] J. S. Chung and A. Zisserman, "Out of time: Automated lip sync in the wild," in *ACCV Workshops*, 2016.

[14] Y. Ding, Y. Xu, S.-X. Zhang, Y. Cong, and L. Wang, "Self-supervised learning for audio-visual speaker diarization," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4367–4371, 2020.

[15] K. Hoover, S. Chaudhuri, C. Pantofaru, I. Sturdy, and M. Slaney, "Using audio-visual information to understand speaker activity: Tracking active speakers on and off screen," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6558–6562.

[16] X. Wang, F. Cheng, G. Bertasius, and D. J. Crandall, "Loconet: Long-short context network for active speaker detection," *ArXiv*, vol. abs/2301.08237, 2023.

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2017.

[18] T. Afouras, J. S. Chung, A. W. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 8717–8727, 2018.

[19] D. Hollosi, J. Schröder, S. Goetze, and J.-E. Appell, "Voice Activity Detection Driven Acoustic Event Classification for Monitoring in Smart Homes," in *3rd Int. Symp. on Applied Sciences in Biomedical and Communication Technologies*, Rome, Italy, Nov. 2010.

[20] S. Wilksen, S. Goetze, D. Hollosi, J.-E. Appell, and J. Bitzer, "Speech Activity Detection for Activity Monitoring using an Embedded Platform," in *Proc. 37th Annual Conference on Acoustics (DAGA)*, Düsseldorf, Germany, Mar. 2011.

[21] G. Datta, T. Etchart, V. Yadav, V. Hedau, P. Natarajan, and S.-F. Chang, "Asd-transformer: Efficient active speaker detection using self and multimodal transformers," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4568–4572.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[23] L. Deininger, B. Stimpel, A. Yüce, S. Abbasi-Sureshjani, S. Schönenberger, P. Ocampo, K. Korski, and F. Gaire, "A comparative study between vision transformers and cnns in digital pathology," *ArXiv*, vol. abs/2206.00389, 2022.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in

2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[25] J. Cartucho, R. Ventura, and M. Veloso, "Robust object recognition through symbiotic deep learning in mobile robots," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 2336–2341.

[26] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," `http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html`.

[27] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 961–970.

[28] J. Liao, H. Duan, K. Feng, W. Zhao, Y. Yang, and L. Chen, "A light weight model for active speaker detection," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 22932–22941.