Neural Computing and Applications (2025) 37:3561–3579
https://doi.org/10.1007/s00521-024-10792-0

ORIGINAL ARTICLE

FabuLight-ASD: unveiling speech activity via body language

Hugo Carneiro[1] · Stefan Wermter[1]

Received: 26 July 2024 / Accepted: 12 November 2024 / Published online: 14 December 2024
© The Author(s) 2024

Abstract
Active speaker detection (ASD) in multimodal environments is crucial for various applications, from video conferencing to human-robot interaction. This paper introduces FabuLight-ASD, an advanced ASD model that integrates facial, audio, and body pose information to enhance detection accuracy and robustness. Our model builds upon the existing Light-ASD framework by incorporating human pose data, represented through skeleton graphs, which minimises computational overhead. Using the Wilder Active Speaker Detection (WASD) dataset, renowned for reliable face and body bounding box annotations, we demonstrate FabuLight-ASD's effectiveness in real-world scenarios. Achieving an overall mean average precision (mAP) of 94.3%, FabuLight-ASD outperforms Light-ASD, which has an overall mAP of 93.7% across various challenging scenarios. The incorporation of body pose information shows a particularly advantageous impact, with notable improvements in mAP observed in scenarios with speech impairment, face occlusion, and human voice background noise. Furthermore, efficiency analysis indicates only a modest increase in parameter count (27.3%) and multiply-accumulate operations (up to 2.4%), underscoring the model's efficiency and feasibility. These findings validate the efficacy of FabuLight-ASD in enhancing ASD performance through the integration of body pose data. FabuLight-ASD's code and model weights are available at https://github.com/knowledgetechnologyuhh/FabuLight-ASD.

Keywords Active speaker detection · Human pose skeleton · Multimodality · Lightweight model

1 Introduction

Active speaker detection (ASD) aims to determine whether a specific person within a video scene is speaking or silent in each frame. This task is essential for various applications, such as speaker diarisation [10, 11], speech enhancement [1], speaker localisation and tracking [6, 27, 28], speech separation [29], and human-robot interaction [35, 36]. By accurately identifying when a person is speaking or silent within a given timeframe, ASD enables the extraction of valuable insights from audiovisual data, driving advancements across multiple domains.

Recent advancements in lightweight and efficient ASD approaches, such as Light-ASD [23], have shown promise for deployment in embedded architectures, such as social robots. These approaches allow fast and accurate determination of active speakers within a group, facilitating more fluid and credible human-robot interactions. However, in scenarios involving large groups or where individuals are at a distance from the robot, the effectiveness of ASD models that rely solely on facial and audio cues is limited. To overcome this challenge, we propose FabuLight-ASD (**F**ace, **a**udio, and **b**ody **u**tilisation for **Light**-weight **A**ctive **S**peaker **D**etection), an extension of Light-ASD that integrates skeleton-based pose information from the target individual. Figure 1 presents FabuLight-ASD's architecture. This enhancement allows the model to complement cues from facial expressions and audio with body pose information, improving its ability to accurately identify active speakers, even in scenarios, where facial nuances are not easily discernible. By leveraging multiple modalities, FabuLight-ASD aims to improve the robustness and accuracy of speaker detection, thereby advancing the

✉ Hugo Carneiro
hugo.carneiro@uni-hamburg.de

Stefan Wermter
stefan.wermter@uni-hamburg.de

[1] Department of Informatics, Knowledge Technology, Universität Hamburg, Vogt-Koelln-Straße 30, 22527 Hamburg, Germany

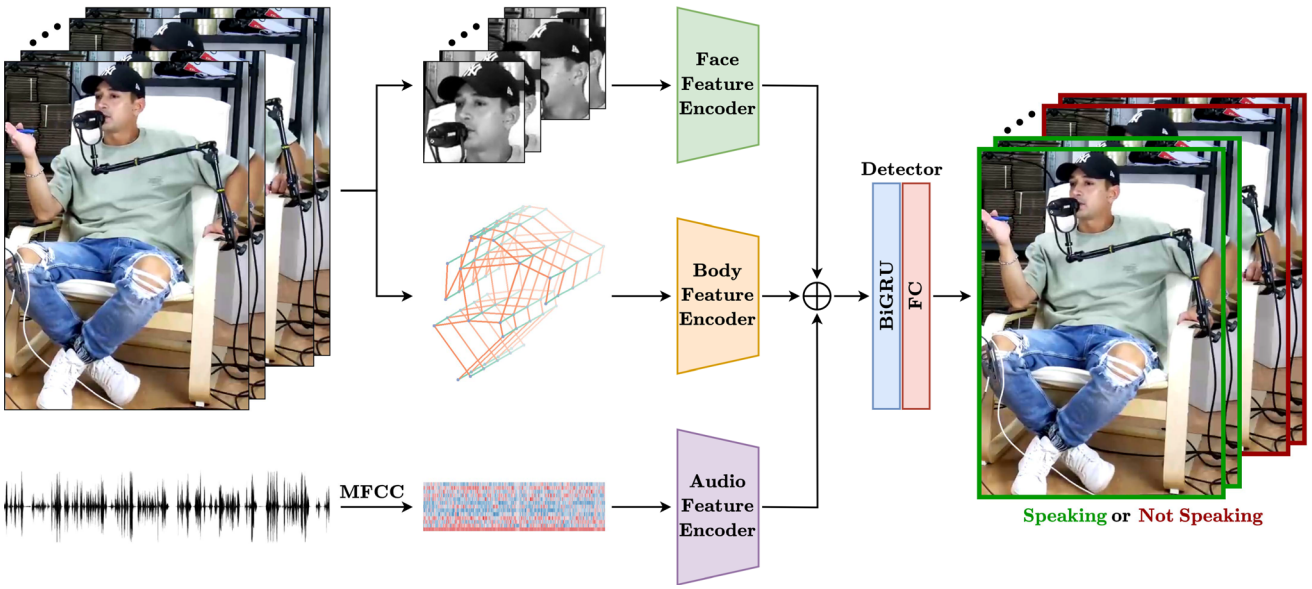


Fig. 1 The architecture of FabuLight-ASD. The model determines whether a target individual is actively speaking or silent in each video frame based on the face crops and body poses of that individual across all frames, along with the corresponding audio information. The overall architecture builds on Light-ASD, with the body feature encoder being a newly added component that improves performance

capabilities of ASD systems for deployment in real-world scenarios, including those involving social robots.

To capitalise on pose information for active speaker detection, we opted to utilise the WASD (Wilder Active Speaker Detection) dataset [32] as a benchmark. WASD stands out for its reliable face position annotations and comprehensive body position annotations. Moreover, WASD presents a diverse range of challenging scenarios, making it conducive to the development of more robust ASD models. Notably, WASD features a high frequency of speaking instances, which is beneficial for applications requiring ASD models that are not biased toward non-speaking individuals [6]. Figure 2 provides examples of the various challenging scenarios in WASD.

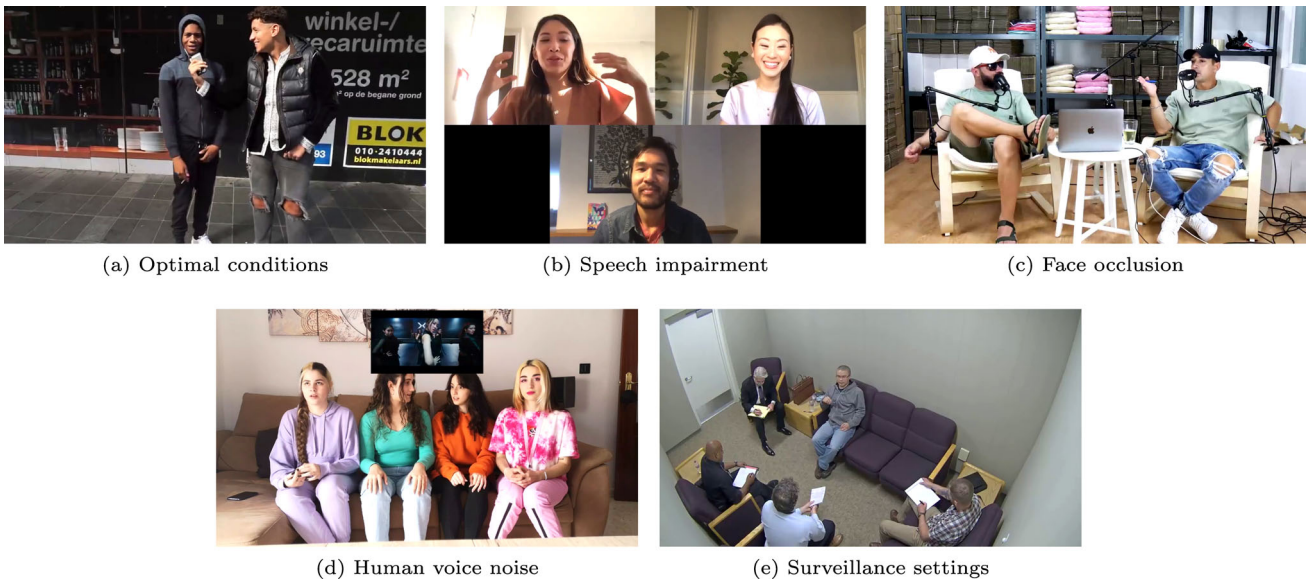
The contributions of this paper are threefold: (1) development of FabuLight-ASD, an ASD model that integrates facial, audio, and pose information to enhance speaker detection performance and robustness; (2) demonstration of the relevance of human pose to detect the source of speaking activity in challenging scenarios; and (3) quantification of FabuLight-ASD's efficiency in terms of model size and computational load.

The structure of the paper is outlined as follows: Sect. 2 provides an overview of the datasets and existing approaches to the active speaker detection task, highlighting their limitations and discussing their potential for extension to incorporate skeleton-based pose information as an additional input modality. In Sect. 3, we examine in detail the architecture of Light-ASD. In Sect. 4, we introduce FabuLight-ASD, our proposed approach, detailing how it integrates human body pose information into the architecture inherited from Light-ASD, and the necessary architectural adjustments for this integration. Section 5 presents the experiments and subsequent analyses conducted to assess the impact of pose information on ASD performance. Section 6 summarises the paper's contributions and suggests directions for future research.

2 Active speaker detection

The task of active speaker detection (ASD) traces back to the pioneering work of Cutler and Davis [12], who employ a time-delayed network to learn audiovisual correlations from speech activity. Historically, ASD solutions relied on small, handcrafted, and task-specific datasets. To address the lack of an in the wild dataset for ASD, Roth et al. [31] introduce AVA-ActiveSpeaker (Atomic Visual Action – ActiveSpeaker), the first large-scale, task-agnostic dataset for active speaker detection. This dataset comprises video footage in various languages and resolutions, depicting individual faces from different angles. It was initially released for the ActivityNet Challenge 2019,[1] where competing models were ranked based on their mean average precision (mAP). Since then, mAP has become the standard metric for comparing ASD models. Each record in AVA-ActiveSpeaker representing an individual in a video frame, assigned a label indicating whether the person is (i) not speaking, (ii) speaking with their voice audible, or (iii) speaking with other audio overlaying their voice.

[1] http://activity-net.org/challenges/2019/tasks/guest_ava.html.



(a) Optimal conditions (b) Speech impairment (c) Face occlusion

(d) Human voice noise (e) Surveillance settings

Fig. 2 Examples of WASD videos of various categories

Since the release of AVA-ActiveSpeaker, numerous approaches have been proposed to tackle in the wild ASD [2–5, 9, 13, 19, 22, 23, 25, 30, 33, 38, 41, 42, 44–47]. However, all of these models have shown significant performance decrease when dealing with small face crops (face width smaller than 64 pixels). In scenarios where face information is unreliable, such as low resolution, occlusion, or when containing non-talking lip movements, body pose information can provide additional cues to disambiguate these cases [32], owing to the correlation between upper-body limb movements and speech activity [16].

Despite being the first large-scale task-agnostic dataset published for active speaker detection, AVA-ActiveSpeaker has several drawbacks. These include a relative scarcity of records featuring speaking activity, several dubbed videos, a lack of support for pose-based active speaker detection, and unreliable face boundary coordinate information. While, AVA-ActiveSpeaker provides annotations for the bounding coordinates of the face crop of a target individual, it lacks corresponding coordinates for the person's body location. Although the body position can be inferred from the scene context and head position information, the reliability of the face bounding box coordinates of AVA-ActiveSpeaker is problematic, posing challenges for accurate body localisation. Specifically, the bounding box coordinates in AVA-ActiveSpeaker are normalised to the scene dimensions, ranging from 0 to 1, based on video dimensions that do not correspond to the actual resolution of the video files in the dataset. Consequently, denormalising the face crop position information can result in bounding boxes that either exclude the target person's face entirely or include it only partially. Figure 3 provides examples of inaccuracies in the head bounding box positions, including cases where the bounding box surrounds something that is not the head of a person, misplacements due to incorrect normalisation or other reasons, and additional bounding boxes nested within each other for the same person.

To address the limitations of AVA-ActiveSpeaker, other datasets have been published, such as Active Speakers in the Wild (ASW) [21] and Wilder Active Speaker Detection (WASD) [32]. WASD, in particular, offers a larger number of records with speaking activity and reliable annotations of both face and body boundary coordinates. Unlike AVA-ActiveSpeaker, both ASW and WASD distinguish only between cases of presence of speaking activity and absence thereof. The correlation between body motion and speech activity has long been studied [39] and has been exploited for active speaker detection in well-controlled datasets [7, 33].

2.1 WASD dataset overview

WASD stands out as the first published dataset for active speaker detection in the wild that contains reliable pose information. In this paper, we employ WASD as our benchmark due to its high diversity and the availability of reliable body boundary coordinate information.

WASD was compiled from 164 YouTube videos from real-world interactions, each capped at a maximum length of 15 min, following a similar practice established during the creation of AVA-ActiveSpeaker. These videos were segmented into clips of up to 30 s, triple the length of the longest clip in AVA-ActiveSpeaker. The dataset comprises a total of 30 h of video annotations, divided into training

(a) Not a person (b) Misplaced bounding box

(c) Incorrect normalised coordinates (d) Additional bounding box

Fig. 3 Examples of bounding box inaccuracies in the AVA-ActiveSpeaker dataset. Correctly placed bounding boxes are indicated in blue and mistake cases in red bounding boxes

and validation sets in an 80/20 proportion, as established by Roxo et al. [32].

The WASD video selection aimed to ensure demographic balance across language, ethnicity, and gender. Additionally, to maintain similar demographic representation, the video quality and face resolution across both training and validation sets, the same clip can be shared by both splits. This approach allows individuals in a scene to be assigned to either split, often resulting in individuals from the same scene being assigned to different splits, which raises a notable concern for models that learn relationships between the target individual and other context individuals present in the scene. For instance, a target individual might be evaluated for their speaking activity in the validation set; however, that same individual could have already been introduced as a context individual during the training of another target individual from the same scene. This overlap can compromise the evaluation, as part of the input for validation may resemble information the model encountered during training. Nevertheless, this aspect of the dataset does not adversely affect models that focus solely on detecting the speaking activity of a target individual without considering context individuals, which is the case for the model proposed in this paper.

WASD videos are categorised into five groups based on the challenge levels of visual and acoustic modalities (cf. Fig. 2). These categories include 'Optimal Conditions' (OC), where there is minimal speech overlap and all speakers are fully visible without occlusion, 'Speech Impairment' (SI), characterised by frequent speech overlap, 'Face Occlusion' (FO), with prevalent cases of mouth occlusion, 'Human Voice Noise' (HVN), featuring frequent overlap of target individuals' voices with background human noise, and 'Surveillance Settings' (SS), comprising camera footage with uncertain face visibility, speech quality, or subject cooperation. Each category is balanced in terms of annotation hours and demographic representation. Roxo et al. [32] propose this categorisation as a means to assess both the adaptability of ASD models to different scenarios and the factors that are more relevant to ASD. We leverage this categorisation to evaluate scenarios where human pose information is most beneficial and those where its utilisation may not yield significant benefits.

2.2 ASD solutions

The publication of the AVA-ActiveSpeaker dataset marked the beginning of research on active speaker detection in the wild. Since then, various approaches have been proposed, including a baseline introduced by the authors of the dataset [31], consisting of a two-stream end-to-end neural network based on the MobileNet architecture [17]. Chung [9] and Zhang et al. [44] devise approaches that surpass this baseline by employing two-stream end-to-end neural networks with 3D convolutions in their visual streams. However, these approaches are limited to short-term windows, roughly 0.5 s long, and rely on large-scale pre-training on lip synchronisation datasets.

To overcome these issues, Alcázar et al. [2] introduce Active Speakers in Context (ASC), an ASD model that employs self-attention to infer inter-speaker relations, allowing for the consideration of long-term contexts, where whole words could be pronounced, approximately 2.25 s. This approach addresses the limitations of large-scale pre-training on lip synchronisation datasets. However, ASC's

architecture is not end-to-end; each speaker's visual and acoustic embeddings have to be provided by a previously trained short-term encoder to capture inter-speaker relationships accurately. After obtaining the embeddings, ASC stacks them as a tensor and passes them through a self-attention layer, which inferred pairwise inter-speaker relations. Subsequently, the tensor output by that layer serves as input to a temporal refinement layer, consisting of a long short-term memory (LSTM). The LSTM acts as a long-term pooling layer, refining the weighted features in the tensor by directly attending to their temporal structure. Notably, ASC was the first openly published ASD model, driving research in the area forward and inspiring direct extensions of the model [5, 22, 34, 46].

Three major enhancements to ASC have greatly improved the performance of ASD models: firstly, the utilisation of graph neural networks for more accurate inference of inter-speaker relations [3, 4, 25]; secondly, the incorporation of self-attention (or transformer-based) layers to enhance the temporal modelling capabilities of the models [19, 38, 45, 47]; thirdly, the adoption of cross-attention mechanisms to capture intermodal signals, thus enhancing the audiovisual synchronisation capabilities of the ASD models and enabling them to correlate the facial movements of a target individual with the scene audio without relying on facial information from other people in the scene [13, 19, 20, 30, 38, 40–42].

It is worth noting that extending these models to include human pose as an additional modality input would have prohibitive impacts for a couple of reasons. Firstly, many models rely on inferring inter-speaker relations for accurate localisation of speaking activity. However, in WASD, individuals in the same scene may be split between training and validation sets. Therefore, to accurately learn inter-speaker relations, data from the validation split could be included as part of the model input during training, contaminating the evaluation process. Secondly, models that leverage visual information exclusively from the target individual utilise cross-attention mechanisms to enhance the correlation between facial movements and scene audio. However, the size of these cross-attention matrices grows quadratically with the length of the scene. Since these matrices correlate modalities pairwise, the number of matrices also increases proportionally to the square of the number of modalities. Given that WASD video blocks can be more than three times longer than the longest video block in AVA-ActiveSpeaker, employing architectures with cross-attention mechanisms becomes impractical, especially for embedded applications. Given these constraints, we chose to build upon Light-ASD [23], the only high-performing ASD model at the time of writing that does not employ cross-attention mechanisms and relies exclusively on visual information from the target speaker to detect speech activity.

3 Light-ASD model architecture

Instead of relying on complex models with high memory and computational requirements, Light-ASD prioritises resource efficiency while maintaining competitive performance. This is accomplished through various strategies, including simplified feature extraction, employing bidirectional gated recurrent units (BiGRUs) [8] for cross-modal modelling, and optimising the model architecture for efficiency [23]. Notably, Light-ASD achieves comparable results to state-of-the-art methods while significantly reducing model parameters and floating-point operations (FLOPs). Its modular design and reduced resource requirements make it easily extendable to incorporate additional modalities and well suited for deployment in resource-constrained environments.

To minimise computational burden, Light-ASD does not leverage the relational contextual information between speakers. Instead, it relies solely on the information of a single target candidate to accurately detect instances of speech activity. Additionally, it is an end-to-end model that does not require pre-training on external training data, resulting in reduced processing time. The model receives a sequence of facial crops of the target individual and the corresponding audio from the video clip as input. The face and audio inputs undergo separate processing by feature encoders specific to each modality, with each encoder producing a feature tensor. These tensors are then fed into a detection module, which assigns a score to each frame, indicating the likelihood of the target individual being actively speaking.

3.1 Feature encoders

Light-ASD's face feature encoder[2] processes 3D stacks of greyscale face images, each with uniform height and width dimensions denoted by N_f. In contrast, the audio feature encoder handles 2D maps consisting of sequences of vectors containing $N_a = 13$ mel-frequency cepstrum coefficients (MFCCs). The length of the image stack represents the number of frames and is denoted by T_f, while that of the MFCC sequence is given by $T_a = 4\,T_f$. To ensure alignment, the raw audio of the video clip is converted to

[2] Liao et al. [23] use the term "visual feature encoder." Here, we replace it with "face feature encoder" to distinguish the processing of facial information from other visual cues, such as body pose. Further terminologies associated with facial information are adjusted accordingly.

MFCCs, and the sequence of MFCC vectors is padded or truncated to match exactly four times the number of frames. Additionally, both face and audio inputs have a channel dimension, denoted as C_{in}, set to 1 for each modality due to the greyscale nature of the images and the single-channel representation of MFCCs. Consequently, the input to the face feature encoder is a tensor of dimensions $1 \times N_f \times N_f \times T_f$, whereas the input to the audio feature encoder is a tensor of dimensions $1 \times N_a \times T_a$. For simplicity, the image height and width dimensions as well as the MFCC vector dimension are hereafter referred to as spatial dimensions, while those associated with the image stack and the MFCC sequence are termed temporal dimensions.

Both feature encoders of Light-ASD share similar architectures, as shown in Figs. 4a and 4b. Each encoder comprises three modality-specific blocks followed by pooling layers. Due to the relation between T_a and T_v, and to ensure identical dimensions for both face and audio features, the first two pooling layers of the audio feature encoder perform a 1D max pool dimensionality reduction on the temporal dimension, while the corresponding layers of the face feature encoder perform a 2D max pool dimensionality reduction operation on the spatial dimensions. Furthermore, in both encoders, the last pooling layer operates on the spatial dimensions; however, a global max pool is performed in the face feature encoder, while a global average pool is utilised in the audio feature encoder.

The architectural similarity also extends to the block level, as shown in Figs. 4c and 4d. Each modality-specific block includes two paths of feature extraction with convolutions of distinct kernel sizes ($\kappa = 3$ and $\kappa = 5$). Instead of using a single high-dimensional convolution that covers all tensor dimensions, the convolution is split into two sequential convolutions, with the first one operating along the spatial dimensions and the next along the temporal dimension. This split in the convolution process significantly reduces the number of parameters in the model. Both spatial and temporal convolutions in either path have the same kernel size. The tensors are padded accordingly to maintain their dimensions after a convolution operation. Batch normalisation [18] and ReLU activation are performed after each convolution. The presence of multiple feature extraction paths ensures a variety of representations, which are then integrated by summing the representations and applying a convolution with a kernel size of 1. All convolutions have a stride of 1 except for the spatial convolutions in both paths of the first block of the face feature encoder, which has a stride of 2. Finally, the number of input and output channels of each block are aligned, with both feature encoders having 1 in-channel and 32 out-channels in their first block, 32 in-channels and 64 out-channels in their second block, and 64 in-channels and 128 out-channels in their last block. The aligned architectures of the feature encoders allow them to output face and audio features Φ_f and Φ_a with identical dimensions, namely $128 \times T_f$.

3.2 Modality fusion and prediction

The fusion of the face and audio features Φ_f and Φ_a is achieved through an element-wise sum. Subsequently, a BiGRU is employed to capture the temporal context inherent in the resulting multimodal representation Φ_{fa}. Leveraging this temporal context, a fully connected (FC) layer provides two scores for every frame of the video clip, indicating the likelihood of the target individual actively speaking or not in that frame. From these scores, Light-ASD performs the prediction using a softmax function.

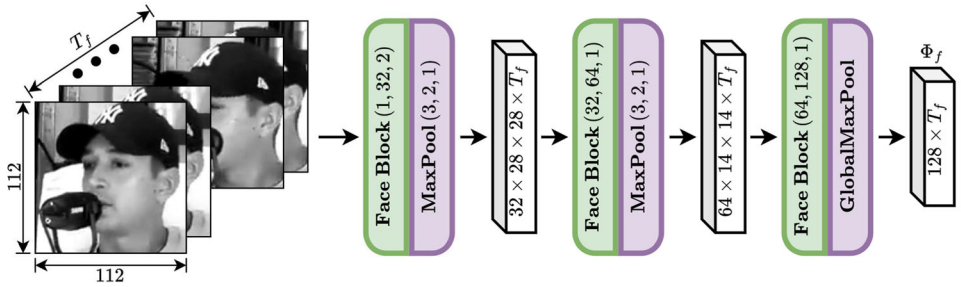
Auxiliary classification heads are also utilised for training purposes in Light-ASD. These heads share a similar structure, comprising a BiGRU and a fully connected layer. However, unlike the main classification head, which integrates all modalities, each auxiliary classification head operates within a purely unimodal framework. Therefore, in each auxiliary classification head, the features output by the feature encoders are directly fed into the corresponding BiGRU.

Light-ASD includes one auxiliary classification head for each auxiliary loss. Notably, while the face auxiliary classifier can determine if a target individual is speaking solely based on facial information, the audio auxiliary classifier can only determine if someone is speaking overall when no facial cues are provided, resulting in high losses. To address this issue, Light-ASD incorporates both the main and face auxiliary losses in its training, excluding the audio auxiliary loss.

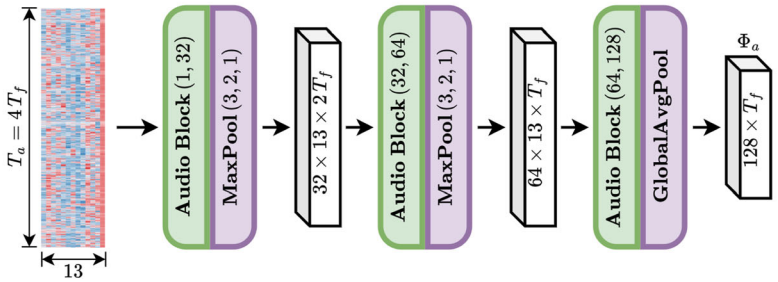
For the computation of the loss function, a temperature parameter τ is utilised to adjust the calculation of the probability of the target individual being actively speaking. The probability of the target individual being actively speaking at the t^{th} frame is determined by

$$p_M^t = \frac{\exp\left(\sigma_{M,spk}^t/\tau\right)}{\exp\left(\sigma_{M,spk}^t/\tau\right) + \exp\left(\sigma_{M,sil}^t/\tau\right)}, \tag{1}$$

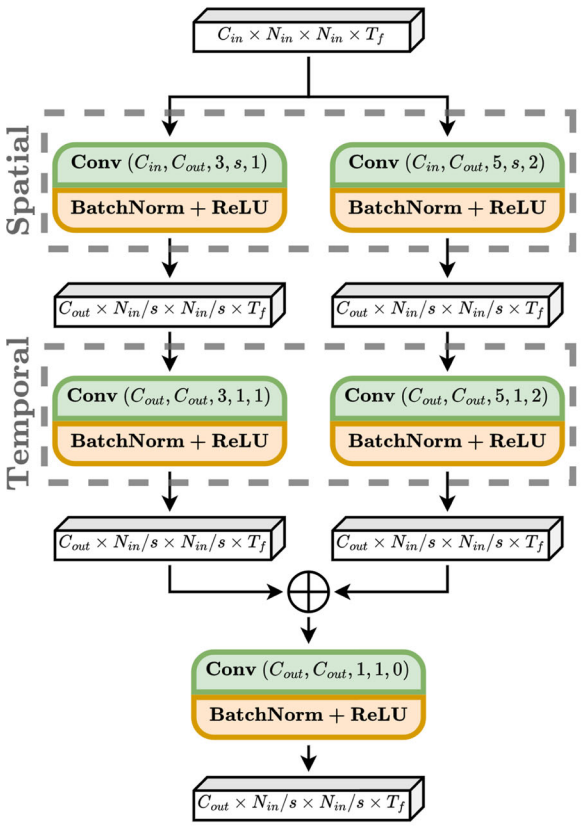
where M represents the modalities used in the classification head, and $\sigma_{M,spk}^i$ and $\sigma_{M,sil}^i$ are the scores assigned by that classification head to the likelihood of the target individual being actively speaking or silent, respectively. The temperature τ progressively decreases with each epoch, inversely correlating with the epoch number, following the formula



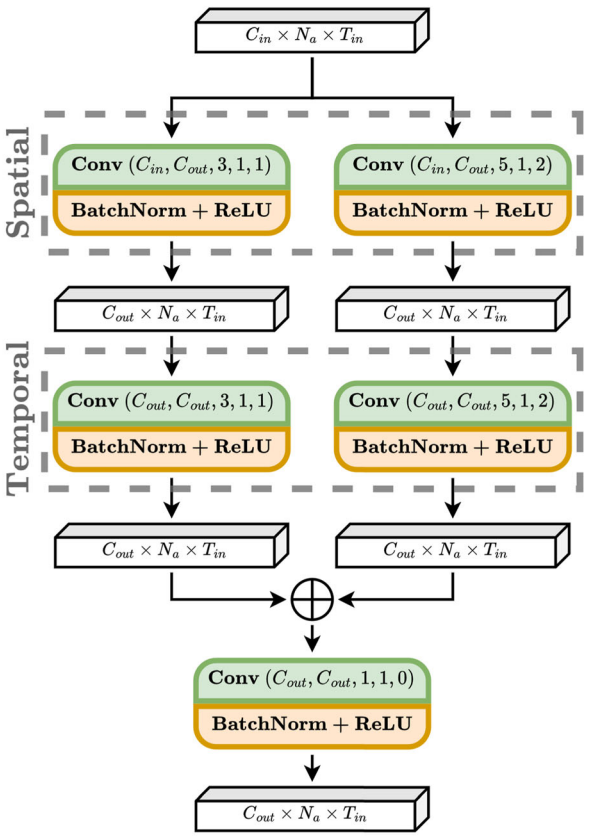
(a) Face feature encoder



(b) Audio feature encoder



(c) Face block



(d) Audio block

◀**Fig. 4** Architectures of Light-ASD feature encoders and their inner blocks. Blocks are given by their number of in-channels C_{in} and out-channels C_{out}, and in the face feature encoder, the stride s in both spatial dimensions. Pooling layer parameters are the kernel size, stride, and padding size. In the face feature encoder, pooling is applied on both spatial dimensions, while in the audio feature encoder, it is applied on the temporal dimension. The parameters of the convolution layers within the blocks are the number of in-channels, number of out-channels, kernel size, stride, and padding size. The labels beside each convolution layer indicate whether the convolution is applied on the spatial or temporal dimensions

$$\tau = 1.3 - 0.02\xi, \tag{2}$$

where ξ is the epoch number. The temperature adjustment facilitates the model's exploration of the solution space, aiding in the avoidance of local optima and guiding the model towards a more refined solution.

Given a video clip with T frames or a batch thereof, one loss $\mathcal{L}_M$ is calculated for each classification head. Light-ASD utilises the main multimodal loss $\mathcal{L}_{fa}$ and the face auxiliary loss $\mathcal{L}_f$. The loss $\mathcal{L}_M$ is a cross-entropy loss, calculated as follows

$$\mathcal{L}_M = \frac{1}{T}\sum_{t=1}^{T}\left(g^t \log\left(p_M^t\right) + (1 - g^t)\log\left(1 - p_M^t\right)\right), \tag{3}$$

where g^t is the ground truth about the status of the target individual as an active speaker at the t^{th} frame, with $g^t = 1$ indicating presence of active speech and $g^t = 0$ indicating absence thereof. The total loss is given by

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{fa} + 0.5\mathcal{L}_f. \tag{4}$$

For evaluation purposes, the temperature is set to a fixed value of $\tau = 1$, and only the probability provided by the multimodal classification is used to determine the presence or absence of active speech. The variation in the value of τ and the probabilities output by the auxiliary classification heads are utilised exclusively during training.

4 FabuLight-ASD model architecture

FabuLight-ASD harnesses body pose information to enhance the inference of whether a target individual is actively speaking. This capability proves particularly valuable in scenarios where the person is distant from the camera, resulting in a low resolution of their facial features.

4.1 Body pose stream

Alongside the face and audio feature encoders, FabuLight-ASD incorporates a body feature encoder. The body pose information is represented as a set of N_b body joints in the COCO format [24].[3] Given the body bounding box coordinates provided in WASD, the HRNet [37] implementation within MMPose [26] is utilised to infer the set of joints corresponding to the person's body delimited by the bounding box coordinates. Each joint is characterised by three values: the horizontal position, the vertical position, and a confidence score in the range $[0, 1]$ indicating HRNet's confidence in the provided positions.

4.1.1 Body feature encoder

Drawing parallels with the architecture of Light-ASD, FabuLight-ASD's body feature encoder incorporates three modality-processing blocks, as shown in Fig. 5. To match the configurations of the face and audio feature encoders inherited from Light-ASD, the number of in-channels and out-channels of each body block in FabuLight-ASD corresponds to those in the other feature encoders, except the in-channels of the first block, which must align with the number of channels of the model input. This input comprises three channels representing each body joint's horizontal position, vertical position, and the confidence of the pose estimation model regarding those coordinates. Accordingly, FabuLight-ASD's body blocks, from the first to the last, have 3 in-channels and 32 out-channels, 32 in-channels and 64 out-channels, and 64 in-channels and 128 out-channels.

To condense the feature representations of each skeleton, a global average pooling operation is applied across the spatial dimension, thus reducing the initial $128 \times N_b$ feature representation to a single 128-dimensional feature vector for each skeleton. Consequently, the body feature encoder processes a sequence of human body pose skeletons with dimensions $3 \times N_b \times T_f$ as input, generating a body feature representation Φ_b whose dimensions align to those of Φ_f and Φ_a, specifically $128 \times T_f$.

4.1.2 Body block architecture

The body blocks within the encoder are based on the spatial-temporal graph convolutional networks (ST-GCNs) proposed by Yan et al. [43]. This choice is motivated by three key factors. Firstly, ST-GCNs excel at capturing dynamic patterns and interactions among body parts during actions. By utilising them as the backbone of its body feature encoder, FabuLight-ASD can discern relationships between body joints, extracting relevant cues about a target individual's speaking activity. Secondly, although initially

[3] We consider two possible ways of feeding the encoder with the information of the body pose of a target individual: (*i*) the set of body joints of the whole body, which encompasses $N_b = 17$ joints in the COCO format; or (*ii*) the set of body joints of the upper body, comprising $N_b = 11$ joints.

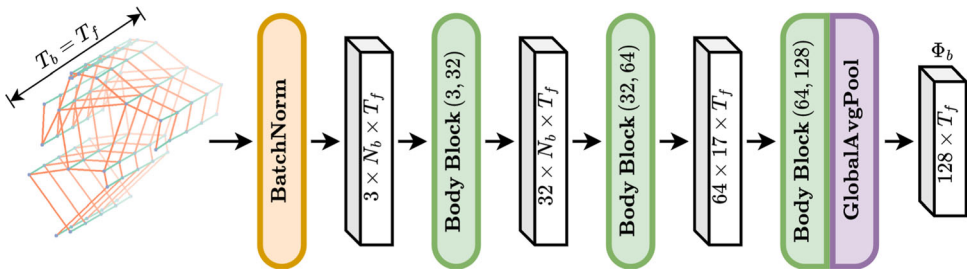


Fig. 5 FabuLight-ASD's body encoder. The parameters of each body block indicate the number of in- and out-channels

designed for action recognition, ST-GCNs have proven versatile and their core concept has been applied in other tasks, such as active speaker detection [3, 4], where dynamic relationships among individuals are represented using graph convolutional networks (GCNs). In these scenarios, individuals are depicted as graph nodes, with edges illustrating their dynamic interactions. Thirdly, ST-GCNs inherently leverage both spatial and temporal information to model the relations between body joints. This aligns with the architecture of the convolutional networks used in the modality-specific blocks of Light-ASD feature encoders, indicating the feasibility of adapting ST-GCNs to maintain the lightweight nature of the architecture while effectively capturing the dynamics of body pose information.

The poses of a person across a sequence of frames are represented as a sequence of skeletons, each with consistently numbered joints. This sequence of skeletons is represented as a graph $G = (V, E)$, where V denotes the joints of the skeleton sequence, with $v_{t,i} \in V$ representing the i^{th} joint at the t^{th} frame. The edge set E comprises two subsets: E_S, depicting joint connections within each frame following the COCO format, and E_F, linking each joint in a frame to corresponding joints in adjacent frames. Formally, $v_{t,i}$ is connected to both $v_{t-1,i}$ and $v_{t+1,i}$. Figure 6 illustrates the spatial-temporal graph of a skeleton sequence.

The body blocks generate increasingly higher-level feature maps on the graph. The convolutions employed in these blocks consider the neighbourhood of each graph node, encompassing intra-body connections and self-connections represented through an adjacency matrix. The concept of node neighbourhood is extended to include nodes connected not only within the same frame, denoted E_S, but also across adjacent frames, denoted E_F. Additionally, the spatial configuration partition strategy divides the spatial neighbour set based on the distance between nodes and their distances to the skeleton's central node $v_{t,c}$ at frame t. This partition distinguishes between concentric and eccentric body part motions. It is represented by $2R + 1$ adjacency matrices $\mathbf{A}^{-R}, \cdots, \mathbf{A}^{R}$, where R is the maximum distance threshold for considering node connections. Specifically, nodes are included in an adjacency matrix if they lie within at most R units from each other. Each adjacency matrix $\mathbf{A}^r = \left[A^r_{ij}\right]_{N_b \times N_b}$ is defined as

$$A^r_{ij} = \begin{cases} 1, & \text{if } r = 0 \text{ and } i = j \\ 1, & \text{if } r < 0,\ d(v_{t,i}, v_{t,j}) = r, \\ & \text{and } d(v_{t,i}, v_{t,c}) \leqslant d(v_{t,j}, v_{t,c}) \\ 1, & \text{if } r > 0,\ d(v_{t,i}, v_{t,j}) = r, \\ & \text{and } d(v_{t,i}, v_{t,c}) > d(v_{t,j}, v_{t,c}) \\ 0, & \text{otherwise} \end{cases}. \tag{5}$$

Here, $d(\cdot, \cdot)$ represents the smallest distance between two nodes. Figure 7 illustrates the labelling of the skeleton graph nodes according to the spatial configuration partition.

Each body block takes as input a tensor $\mathbf{X} \in \mathbb{R}^{C_{in} \times N_b \times T_{in}}$ and produces a tensor $\mathbf{Y} \in \mathbb{R}^{C_{out} \times N_b \times T_{out}}$. Each block is associated with a spatial kernel size κ_S and a temporal kernel size κ_T. The former is set at $\kappa_S = 2R + 1$, representing the number of spatial configuration partitions, while the latter, although unrelated to the adjacency matrices, is set to be identical to κ_S. Akin to Light-ASD's modality-specific blocks, FabuLight-ASD's body blocks incorporate two paths of feature extraction: one utilising kernel sizes $\kappa_S = \kappa_T = 3$, and the other with $\kappa_S = \kappa_T = 5$. The determination of κ_S for each path involves setting $R = 1$ for the former and $R = 2$ for the latter, guided by the relation $\kappa_S = 2R + 1$, which defines the spatial kernel size based on the number of spatial configuration partitions.

Both paths comprise a graph convolution followed by a temporal convolution, as shown in Fig. 8. The graph convolution over an input $\mathbf{X} \in \mathbb{R}^{C_{in} \times N_b \times T_{in}}$ occurs in two stages: Firstly, a 2D convolution yields $\mathbf{M} \in \mathbb{R}^{\kappa_S \times C_{out} \times N_b \times T_{in}}$ using a learnable weight tensor $\mathbf{W}$ with a kernel of size 1. Next, via a tensor contraction operation, the spatial-temporal graph convolution generates a feature tensor $\mathbf{Z} \in \mathbb{R}^{C_{out} \times N_b \times T_{in}}$ whose channel-wise slices $\mathbf{Z}_c \in \mathbb{R}^{N_b \times T_{in}}$ are given by

$$\mathbf{Z}_c = \sum_{r=-R}^{R} (\mathbf{B}^r)^\top \mathbf{M}^r_c, \tag{6}$$

where $\mathbf{M}^r_c \in \mathbb{R}^{N_b \times T_{in}}$ represents a slice of $\mathbf{M}$, and $\mathbf{B}^r \in$

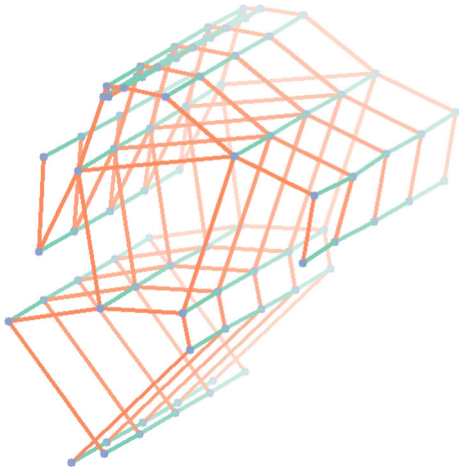


Fig. 6 Spatial-temporal graph representing a sequence of body poses. Body joints are marked in blue. Edges indicating joint connections within the same frame are coloured in orange. Temporal connections, which link each joint in a given frame to the corresponding joints in adjacent frames, are represented in green. Differing alpha channels are used to indicate body pose skeletons in different frames and for visualisation purposes

$\mathbb{R}^{N_b \times N_b}$ is a learnable matrix, which is initialised according to

$$\mathbf{B}^r = (\mathbf{D}^r)^{-\frac{1}{2}} \mathbf{A}^r (\mathbf{D}^r)^{-\frac{1}{2}}, , \tag{7}$$

where $\mathbf{A}^r$ is the adjacency matrix defined by Eq. 5, and $\mathbf{D}^r$ is a diagonal normalisation matrix, with $D^r_{ii} = \sum_k (A^r_{ik}) + \varepsilon$. Here, ε is set to 0.001 to prevent empty rows in $\mathbf{D}^r$. Subsequently, the temporal convolution entails a 2D convolution with kernel dimension $\kappa_T \times 1$ over $\mathbf{Z}$, resulting in a feature representation $\mathbf{Y} \in \mathbb{R}^{C_{out} \times N_b \times T_{out}}$. The feature representations produced by both paths are then integrated via summation followed by a convolution with a kernel size of 1. Notice that each learnable parameter described in this paragraph has a copy in every path of all body blocks of FabuLight-ASD, which are initialised identically but fine-tuned differently.

Batch normalisation is applied after each convolution operation and as the initial operation on the input data, before it passes through the first body block. The ReLU activation function is applied after the batch normalisation following the graph convolution and at the end of each block.

4.2 Modality fusion and prediction

Similarly to Light-ASD, FabuLight-ASD features one main classification head and some auxiliary classification heads, which are exclusively used for training purposes. In Light-ASD, the main classification head generates scores based on the multimodal representation obtained by fusing the

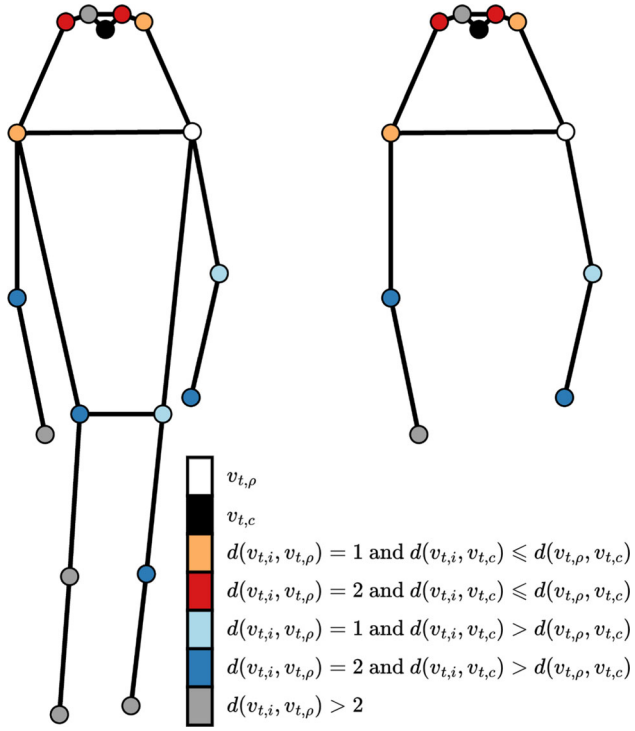


Fig. 7 Spatial-configuration-based node partition of the body pose skeleton graph using the COCO pose template. The left side illustrates the partitioning for the whole body, while the right side focuses on the upper body. For the COCO template, ST-GCN treats the nose node as the central node $v_{t,c}$ for determining the spatial configuration. The spatial configuration is the same for every time frame t. In this example, the right shoulder node is taken as the root node $v_{t,\rho}$. Nodes $v_{t,i}$ are marked based on their distance from the root node and whether they lie closer or further from $v_{t,c}$ than the root node. This example represents the case in which $R = 2$ and the spatial kernel size is $\kappa_S = 5$. Given this configuration, the adjacency matrix is determined according to Eq. 5

features produced by both feature encoders, namely Φ_f and Φ_a. Additionally, Light-ASD has a single auxiliary classification head, responsible for generating scores solely based on the face feature Φ_f. In contrast, FabuLight-ASD processes three input modalities: face (as a sequence of face crops of the target individual), audio (as an MFCC tensor derived from the audio clip), and body pose (as a spatio-temporal graph of the body joints of the target individual), allowing fusion by combining them. The main classification head of FabuLight-ASD generates scores based on the multimodal representation provided by fusing all modality features, namely Φ_f, Φ_a, and Φ_b. Furthermore, FabuLight-ASD includes two auxiliary classification heads that utilise unimodal features, specifically face and body. FabuLight-ASD lacks an audio auxiliary classification head for the same reason as Light-ASD, as discussed in Sect. 3.2.

Given a set of modalities M, a classification head provides two scores, $\sigma^t_{M,spk}$ and $\sigma^t_{M,sil}$, for every frame t of an

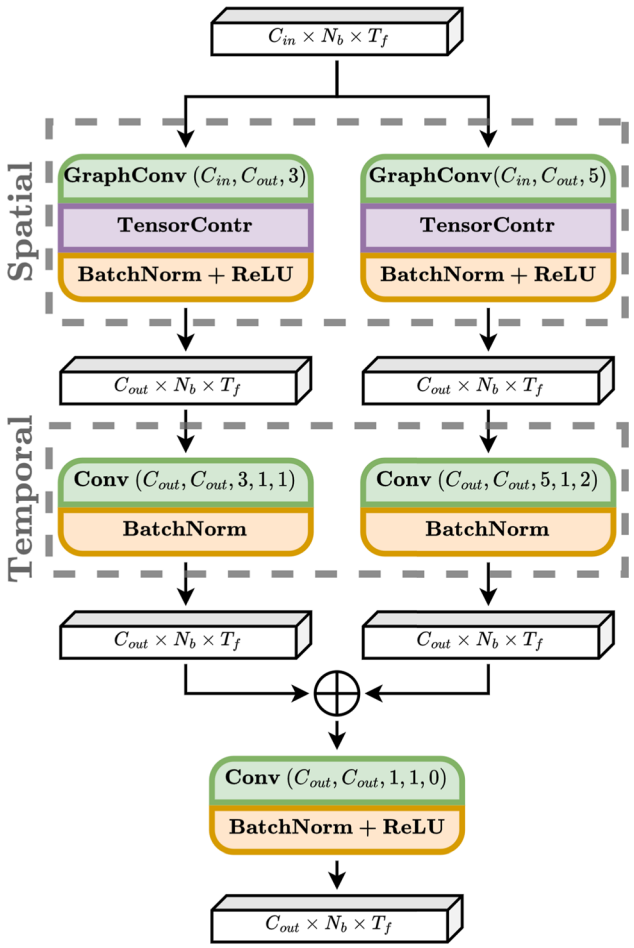


Fig. 8 Architecture of FabuLight-ASD's body block. The parameters of each component follow the description given in Fig. 4. TensorContr stands for tensor contraction, representing the operation described by Eq. 6

input video footage, indicating the likelihood of a target individual being actively speaking or not in each frame respectively. Akin to Light-ASD's classification procedure, unimodal features are summed element-wise. Subsequently, a BiGRU is applied to the resulting sum, followed by an FC layer that outputs the scores. Both the main classification head and the auxiliary ones can predict the speaking activity of an individual across a sequence of frames using Eq. 1. The predictions of all classification heads and the temperature parameter τ are used exclusively during training. However, during evaluation, only the prediction obtained from the main classification head is employed, and no temperature adjustment is applied.

Given the probabilities p^t_{fab}, p^t_f, and p^t_b, computed from the scores output by the classification heads, the corresponding cross-entropy losses are calculated using Eq. 3. The total loss of the model is then determined by

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{fab} + 0.25\mathcal{L}_f + 0.25\mathcal{L}_b. \tag{8}$$

5 Experiments and analyses

In this section, we present a detailed account of the experiments conducted to evaluate the performance of FabuLight-ASD. We first outline the implementation details and then compare FabuLight-ASD to the baseline Light-ASD model using the WASD dataset. Next, we provide a detailed performance breakdown to understand the impact of various factors such as face resolution, pose estimation confidence, and the number of speakers on the model's performance. Finally, we analyse the efficiency of the models in terms of parameter count and multiply-accumulate (MAC) operations, demonstrating that the additional computational cost is negligible.

5.1 Implementation details

For the following experiments, we utilised the default parameters of Light-ASD. Namely, batches were built by combining videos with the same number of frames as long as the total number of their frames did not surpass 2000, training was performed until a maximum of 30 epochs, FabuLight-ASD was optimised with ADAM with an initial learning rate of 10^{-3} and learning decay rate of 0.05 per epoch.

5.2 Evaluation of fabuLight-ASD

The results shown in Table 1 provide a comparative analysis between the performances of two variants of FabuLight-ASD – one that utilises information from the whole body of a target individual and one that employs information only from their upper body – and ASD models evaluated by Roxo et al. [32] on the WASD dataset. Among these models, Roxo et al. [32] include two adaptations of TalkNet, denoted as TalkNet (body + audio) and TalkNet (face + body + audio). The former adaptation receives as input a stack of greyscale body crop images and the audio of the video clip but no face crops, whereas the latter uses input data from all three modalities. It is worth noting that Roxo et al. [32] do not provide the overall mean average precision (mAP) performance of those adaptations in their paper.

We evaluated the performance of both FabuLight-ASD variants on the WASD dataset in terms of mAP across different categories. Notably, both variants achieved higher overall performance than every other model, indicating the benefit of incorporating body pose information into the task of determining whether a given person is the source of some perceived speech activity. Furthermore, the upper-

Table 1 Comparison of performance on the WASD dataset for various ASD models, as obtained by Roxo et al. [32], and for two variants of FabuLight-ASD. Values represent the mean average precision (mAP) of each model across WASD categories: OC (optimal conditions), SI (speech impairment), FO (face occlusion), HVN (human voice noise), and SS (surveillance settings). Citations beside model names refer to the papers, where the models were introduced. All results were obtained by Roxo et al. [32]. The highest mAPs within each category are marked in bold. Italicised values indicate mAPs higher than those of both FabuLight-ASD variants

Model	OC	SI	FO	HVN	SS	Overall
ASC [2]	91.2	92.3	87.1	66.8	72.2	85.7
MAAS [3]	90.7	92.6	87.0	67.0	76.5	86.4
ASDNet [22]	96.5	97.4	92.1	77.4	*77.8*	92.0
TalkNet (face + audio) [38]	95.8	97.5	93.1	81.4	*77.5*	92.3
TS-TalkNet [19]	96.8	97.9	94.4	84.0	*79.3*	93.1
Light-ASD [23]	97.8	98.3	95.4	84.7	*77.9*	93.7
TalkNet (body + audio) [32]	91.1	95.5	88.4	73.1	75.0	—
TalkNet (face + body + audio) [32]	96.9	98.1	95.4	83.8	***81.5***	—
FabuLight-ASD (upper body)	97.7	**98.6**	**96.1**	**86.4**	77.3	**94.3**
FabuLight-ASD (whole body)	**98.1**	**98.6**	96.0	85.6	77.1	94.0

body variant presents a slight enhancement in overall performance compared to the whole-body variant. FabuLight-ASD outperforms every other model in each video category, except for surveillance settings (SS), which are the most challenging scenarios in WASD. In this category, TalkNet (face + body + audio) achieves the highest performance with 81.5% mAP, compared to 77.3% and 77.1% for the upper-body and whole-body variants of FabuLight-ASD, respectively. This high performance in surveillance settings, however, occurs at the expense of the model's performance in the remaining categories. Additionally, both TalkNet adaptations are very inefficient compared to FabuLight-ASD, as they include additional cross-attention modules that considerably increase the number of multiply-accumulate (MAC) operations and the models' number of parameters.

Overall, results indicate the benefit of body pose information across all video categories. The noticeable improvements shown by FabuLight-ASD in conditions with speech impairment (SI), face occlusion (FO), and human voice noise (HVN) suggest that the inclusion of body pose information as a spatial-temporal body pose graph is particularly useful in these scenarios. While stacks of greyscale body crop images improve performance in surveillance settings, they do so at the expense of performance in other video categories. Moreover, the use of spatial-temporal body pose graphs not only enhances performance in most scenarios but also keeps the model lightweight, highlighting the advantages of FabuLight-ASD.

5.3 Performance breakdown

To gain deeper insights into FabuLight-ASD's performance compared to Light-ASD,[4] we divided the validation set of WASD into mutually exclusive groups based on key factors: face resolution, pose estimation confidence, the number of individuals in a scene, and the temporal span of the inputs. By examining the mAP across these divisions, we aim to identify how variations in these conditions impact active speaker detection. This approach allows us to better understand the strengths and limitations of FabuLight-ASD in comparison to Light-ASD, and to evaluate the effectiveness of incorporating body pose information. Table 2 summarises the performance of Light-ASD and FabuLight-ASD across various subsets of WASD according to the specific ablation criteria detailed in Sects. 5.3.1 to 5.3.4.

It is important to note that the ablation criteria might affect videos from different categories in varied ways. Some categories might contain videos with a higher number of individuals that align more with one criterion than another. This results in a different number of samples being evaluated category-wise, which in turn can lead to the overall mAP metric becoming more biased towards videos from the category more represented by a given ablation criterion. Additionally, some ablation criteria may not apply to certain categories. For example, there are no samples in the subset of "speech impairment" videos, where an individual received a high average confidence

[4] Despite TalkNet (face + body + audio) achieving superior performance in surveillance settings, Light-ASD achieved the highest overall mAP among the models evaluated by Roxo et al. [32]. Light-ASD was deliberately included in this section due to its similar architecture to FabuLight-ASD, facilitating a comparative analysis and providing insights into the impact of body pose information across different subsets of the WASD dataset.

Table 2 Performance breakdown across various conditions, where "None" in the Body column corresponds to Light-ASD, and "Upper" and "Whole" correspond to the two variants of FabuLight-ASD. The bold indicates the highest mAP within each category for each performance breakdown condition

Ablation	Body	OC	SI	FO	HVN	SS	Overall
Face resolution							
Small	None	96.9	**99.9**	**93.7**	48.1	**77.7**	**81.7**
	Upper	**97.3**	99.8	92.2	**60.5**	75.5	80.2
	Whole	**97.3**	**99.9**	92.0	59.7	77.2	**81.7**
Medium	None	97.9	96.7	93.4	86.8	78.4	93.0
	Upper	**98.0**	97.2	95.2	88.0	**79.7**	93.7
	Whole	**98.0**	**97.4**	**95.5**	**88.6**	77.6	**93.9**
Large	None	97.8	98.9	**97.5**	84.4	**77.3**	97.2
	Upper	**98.3**	**99.3**	97.3	84.6	73.2	**97.4**
	Whole	97.6	99.2	97.3	**85.5**	72.2	97.3
Upper-body pose estimation confidence							
Low	None	**88.3**	**97.5**	**97.0**	**79.6**	64.1	82.8
	Upper	87.9	96.4	96.4	76.0	**68.8**	**83.8**
	Whole	86.5	96.5	95.4	77.6	56.6	80.3
Medium	None	95.5	98.5	96.2	86.1	72.9	93.4
	Upper	**96.8**	**98.8**	**96.4**	86.5	**73.2**	93.8
	Whole	95.2	**98.8**	96.1	**87.1**	72.6	**93.9**
High	None	98.5	97.2	94.8	83.2	**84.5**	94.3
	Upper	**98.6**	**97.9**	95.7	85.1	82.1	94.6
	Whole	98.5	**97.9**	**96.1**	**86.0**	83.8	**95.2**
Whole-body pose estimation confidence							
Low	None	95.5	98.7	97.6	84.4	71.4	95.6
	Upper	**96.9**	**98.9**	**97.9**	**85.0**	**74.2**	**96.1**
	Whole	94.7	**98.9**	97.4	**85.0**	71.3	95.8
Medium	None	98.1	97.3	94.6	85.3	**79.6**	92.8
	Upper	**98.4**	**97.9**	95.4	86.2	77.4	93.0
	Whole	98.2	**97.9**	**95.7**	**87.5**	78.8	**93.6**
High	None	**98.1**	—	**95.5**	80.0	77.5	92.4
	Upper	97.4	—	95.0	81.7	**79.7**	92.4
	Whole	97.7	—	95.4	**82.5**	77.4	**92.8**
Number of speakers							
Two	None	98.4	98.7	97.4	86.3	**83.8**	95.6
	Upper	**98.6**	99.0	**97.8**	86.6	**83.8**	**96.1**
	Whole	98.5	**99.1**	97.6	**87.7**	83.0	96.0
Three	None	92.9	98.9	93.2	77.8	**74.2**	90.3
	Upper	**93.4**	**99.1**	93.6	**81.3**	72.4	90.1
	Whole	91.4	**99.1**	**94.5**	81.2	73.7	**91.1**
At least four	None	98.2	96.8	89.4	85.0	51.8	92.7
	Upper	**98.8**	**97.6**	90.9	86.1	53.3	93.7
	Whole	98.2	97.5	**92.1**	**87.4**	**54.4**	**93.9**
Input temporal span							
Short	None	95.6	**99.5**	—	85.2	**81.8**	88.9
	Upper	97.0	99.0	—	83.6	77.8	88.9
	Whole	**97.8**	98.8	—	**86.5**	80.2	**90.6**

Table 2 (continued)

Ablation	Body	OC	SI	FO	HVN	SS	Overall
Medium	None	**98.5**	**97.6**	95.1	82.1	72.1	94.2
	Upper	98.3	97.5	**96.0**	**85.2**	**82.5**	**95.1**
	Whole	97.5	97.2	95.3	84.1	61.7	93.6
Long	None	97.7	98.3	95.5	85.1	**78.3**	93.7
	Upper	97.6	**98.7**	**96.1**	**86.7**	77.4	**94.3**
	Whole	**98.1**	**98.7**	**96.1**	85.9	78.0	94.1

score for the estimation of their whole-body joints (cf. Table 2).

5.3.1 Face resolution

We divide the data into three groups based on face resolution: large faces (widths greater than 128 pixels), middle faces (widths between 64 and 128 pixels), and small faces (widths smaller than 64 pixels). This division helps to understand the impact of face image size on the model's performance. The category-wise mAP is evaluated for each group, allowing us to see how the model handles different face resolutions under varying conditions.

The results in Table 2 indicate a clear benefit in utilising body pose information in scenarios with human voice background noise, especially when the resolution of the face of the target individual is rather small. Light-ASD presents higher performance for videos with face occlusion and in surveillance settings, yet this is somewhat erratic, as its performance is higher when face resolutions are either small or large. On the other hand, when the face of the target individual presents a medium size, it seems that the pose information has a positive impact, especially that of the upper body. Moreover, in the less challenging scenarios (OC and SI), the upper-body version of FabuLight-ASD slightly outperforms Light-ASD regardless of the face resolution.

In addition, when comparing the upper-body and whole-body variants of FabuLight-ASD, specific trends emerge regarding face resolution. For small face resolutions, both variants perform similarly across most categories, except in HVN, where the upper-body variant shows superior classification accuracy. However, in SS, the whole-body variant performs better, resulting in its overall superiority in small-resolution samples.

For medium or large face resolutions, the trends shift. The whole-body variant performs better in the HVN category, while the upper-body variant excels in SS. In the OC category, both variants perform similarly when the face resolution is medium, but the upper-body variant gains an advantage when the face resolution is large. Overall, both variants perform comparably for medium and large face resolutions.

5.3.2 Pose estimation confidence

For pose estimation confidence, we evaluate the model's performance using two sets of body joints: the upper-body joints and the whole-body joints. Each set is divided based on the average confidence score of the joints: high confidence (average confidence score of the selected set of body joints greater than 0.75), medium confidence (average confidence score between 0.5 and 0.75), and low confidence (average confidence score lower than 0.5). This evaluation reveals how the reliability of pose information affects detection accuracy in different WASD categories.

As expected, no significant positive impact has been found by utilising human body pose information with a low average confidence score. However, while this generally leads to a lower performance for both variants of Fabu-Light-ASD compared to Light-ASD, the upper-body variant performs comparatively better in surveillance settings (cf. Table 2). This result may be influenced by the characteristics of the videos in this category. It is also important to note that the surveillance settings category contains a disproportionately higher number of samples with low-confidence body pose estimations compared to other categories. This higher sample size in surveillance settings likely contributes to the upper-body variant's overall higher mAP. In other words, the better performance of the upper-body variant in this category biases its overall results. Thus, while low-confidence pose information does not generally improve performance, its impact on the upper-body variant seems less detrimental in specific contexts like surveillance settings, where pose data might still provide useful contextual cues despite lower confidence scores.

Conversely, the results in Table 2 show that the utilisation of medium-confidence and high-confidence upper-body poses is beneficial in determining whether a target individual is the source of some speech activity. Furthermore, although the utilisation of whole-body pose information is not promising when there is medium average

confidence in the body joints of the upper body, the whole-body pose information presents a positive impact in active speaker detection if the upper-body joints of the target individual present a high confidence score on average. Counterintuitively, in surveillance settings, Light-ASD outperforms both variants of FabuLight-ASD when there is an average high confidence in the upper-body joints of the target individual. However, this is not reflected in the overall mAP, which indicates that this high performance of Light-ASD might be due to an exceptional case.

Finally, no useful analysis can be derived from the performance breakdown based on whole-body joint average confidence scores. Given how this criterion is calculated, several cases categorised as low average confidence have been categorised as medium or high average confidence according to the upper-body pose ablation criterion. Consequently, the results for the low-confidence whole-body pose criterion tend to be close to those of the medium-confidence upper-body pose criterion, and to a lesser extent to those of the high-confidence upper-body pose criterion. Similarly, the results for the medium-confidence whole-body pose criterion closely resemble those of the cases that fit into the high-confidence upper-body pose criterion. Furthermore, the number of dataset samples that fit the high-confidence whole-body pose criterion is rather small (cf. the lack of SI samples that fit this criterion), leading to less informative results.

When comparing the FabuLight-ASD variants based on pose estimation confidence, a few key patterns emerge. For samples with low upper-body confidence, both variants underperform relative to Light-ASD, though the upper-body variant excels in SS and generally outperforms the whole-body variant across most categories, except in HVN. At medium confidence levels, both variants perform similarly overall. Still, the upper-body variant shows clear advantages in OC, FO, and SS. In contrast, the whole-body variant performs better in HVN when whole-body confidence is low.

The whole-body variant proves more effective at high upper-body confidence, especially in challenging categories like FO, HVN, and SS, consistently outperforming the upper-body variant. While the whole-body variant also performs better in the small subset of samples with high whole-body confidence, particularly in HVN, the upper-body variant remains superior in SS. Overall, the whole-body variant tends to perform better as pose confidence increases, particularly in more challenging scenarios.

5.3.3 Number of speakers

We divide the WASD validation set based on the number of individuals present in a scene: two individuals, three individuals, and four or more individuals. This analysis reveals how the complexity of the scene, in terms of the number of speakers, influences the model's ability to detect active speakers accurately. The category-wise mAP is calculated for each group to understand the performance variations across different scenarios.

Results in Table 2 reveal an improvement in active speaker detection when either upper- or whole-body pose is employed in non-surveillance videos. Specifically, there is a slight improvement in less challenging scenarios (OC and SI) and a much higher improvement in scenarios with face occlusion and human voice background noise. Although Light-ASD performs better in surveillance settings, this is particularly true in videos with three speakers, but the performance difference is noticeably small in videos with only two speakers. Finally, in videos with four or more speakers, both variants of FabuLight-ASD outperform Light-ASD in every video category, especially in the most challenging scenarios.

For scenes with two speakers, both FabuLight-ASD variants perform similarly in OC, SI, and FO categories. However, the whole-body variant outperforms in the HVN category, while the upper-body variant excels in SS. Overall, both variants perform comparably in two-speaker scenes.

In scenes with three speakers, the upper-body variant has an advantage in OC samples, but the whole-body variant surpasses it in both FO and SS, leading to better overall performance. In scenes with four or more speakers, the whole-body variant consistently outperforms the upper-body variant, particularly in more challenging categories like FO, HVN, and SS, though the upper-body variant retains an edge in OC samples.

5.3.4 Temporal span

To examine the effect of input temporal span on model performance, we categorised the WASD validation set into three groups: short (up to 300 frames), medium (301 to 600 frames), and long (more than 600 frames). The temporal span, represented as T_f in Sect. 3.1, corresponds to the size of the input modalities for both FabuLight-ASD and Light-ASD. By evaluating the category-wise mAP for each group, we can assess how varying input durations affect the performance of both models.

The performance breakdown in Table 2 shows that for short temporal spans, whole-body pose data provides greater advantages than upper-body data, particularly in OC and HVN, where the whole-body version of FabuLight-ASD performs best. For medium temporal spans, the upper-body version of FabuLight-ASD outperforms both the whole-body variant and Light-ASD, especially in challenging categories like FO, HVN, and SS. It achieves

Table 3 Impact of inclusion of body pose information as an input modality on model efficiency

Model	Params (M)	MACs (M)
Light-ASD	1,021	204
FabuLight-ASD (upper)	1,300	207
FabuLight-ASD (whole)	1,300	209

the highest performance overall across all temporal span groups. For long temporal spans, both FabuLight-ASD variants either outperform or remain competitive with Light-ASD in every category except SS, where Light-ASD holds a slight advantage. Nevertheless, the upper-body version of FabuLight-ASD achieves the highest overall performance for long temporal spans.

These findings indicate that whole-body pose data is more effective for shorter sequences, while upper-body pose data becomes increasingly valuable in medium and long temporal spans, particularly in challenging categories like FO, HVN, and SS. The upper-body version of FabuLight-ASD performs best in medium temporal spans and maintains strong performance in long spans, especially when compared to Light-ASD.

For short temporal spans, the upper-body variant outperforms the whole-body variant in key categories, including OC, HVN, and SS, resulting in better overall performance. For medium temporal spans, the upper-body variant consistently surpasses the whole-body variant across all categories and overall. In long temporal spans, the upper-body variant continues to excel in HVN samples, while the whole-body variant performs better in OC and SS. Overall, both variants perform similarly for long temporal spans.

5.4 Model efficiency

Light-ASD stands out as a high-performing ASD model known for being efficient both spatially (retaining a small number of parameters – 1,021 million) and temporally (executing a low number of operations – 204 million MACs[5] per input frame). Table 3 demonstrates that the inclusion of a pose feature encoder increases the parameter count to 1,300 million, marking a 27.3% rise solely attributable to the body feature encoder. Notably, the substitution in the detector component has not altered the parameter count. In other words, the addition of one input modality showed no impact on the parameters of the multimodal fusion mechanism, which comprises a bidirectional GRU and a fully connected layer. The increase in parameters is identical in both variants of FabuLight-ASD due to identical architectures.

The efficiency of both FabuLight-ASD variants is highlighted not only by the modest rise in parameter count compared to Light-ASD but also by the negligible growth in MAC operations. Table 3 indicates that incorporating and processing a sequence of whole-body skeleton graphs results in a 2.4% increase in MAC operations. Furthermore, this increase diminishes to 1.5% when using information solely from the upper body. These slight increments in MAC operations primarily stem from the vast majority of operations being performed in the convolutions within the face feature encoder. This underscores the benefits of using pose skeleton graphs as a representation of the body pose modality instead of sequences of whole-body images of the target individual, which would significantly increase the number of MAC operations executed by the model. It is important to note that these efficiency comparisons are based solely on the model architecture as evaluated by Liao et al. [23], and do not include additional preprocessing steps such as body pose estimation or human face detection, which could impact overall computational efficiency in real-world applications. Nevertheless, the inherently low number of parameters and computational cost of FabuLight-ASD may offset the effects of these additional steps.

6 Conclusion

This study underscores the potential of integrating multiple modalities including body pose information to improve the accuracy and robustness of active speaker detection (ASD) models. We have introduced FabuLight-ASD, a lightweight model that combines facial, audio, and body pose data to achieve improved speaker detection performance. Through extensive experiments on the WASD dataset, we demonstrate the effectiveness of incorporating body pose information, particularly in challenging scenarios like multiple speakers or face occlusion. FabuLight-ASD outperforms other ASD models, especially in categories such as face occlusion and human voice background noise, highlighting the significance of body pose data for handling complex conditions, while maintaining computational efficiency.

Our research paves the way for future investigations in several promising directions. These include exploring real-time developments of FabuLight-ASD in social robots to validate its utility in diverse environments, particularly in multiparty human-robot interaction (MHRI) scenarios, where the system must identify active speakers among multiple participants. The utilisation of datasets with egocentric perception videos that include active speaker detection tasks [14, 15] would also be beneficial for further

[5] This value was calculated with the 'calflops' Python package found at https://pypi.org/project/calflops/.

testing and enhancing the model's performance. Furthermore, leveraging body pose information to reduce reliance on face information could allow for smaller face dimensions as inputs to the face feature encoder, thereby decreasing the model's MAC operations and improving efficiency. Finally, implementing adaptive weighting techniques to account for the varying contributions of each modality based on their quality could further enhance the model's robustness. These avenues of future work will not only refine FabuLight-ASD but also contribute to the broader field of active speaker detection, ultimately enabling more effective and adaptive human-robot interactions.

Acknowledgements The authors thank Cornelius Weber for his valuable feedback on this paper and acknowledge partial support from the German Research Foundation DFG under project CML (TRR 169).

Author contributions Conceptualisation was done by Hugo Carneiro and Stefan Wermter; Methodology was done by Hugo Carneiro; Formal analysis and investigation were done by Hugo Carneiro; Writing - original draft preparation were done by Hugo Carneiro; Writing - review and editing were done by Stefan Wermter; Funding acquisition was done by Stefan Wermter; Resources was done by Stefan Wermter; Supervision was done by Stefan Wermter.

Funding Open Access funding enabled and organized by Projekt DEAL. This work is partially supported by the German Research Foundation DFG under project CML (TRR 169).

Availability of data and materials This work makes use of publicly available data.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Code Availability FabuLight-ASD's code and model weights are available at https://github.com/knowledgetechnologyuhh/FabuLight-ASD.

References

1. Afouras T, Chung JS, Zisserman A (2018) The conversation: Deep audio-visual speech enhancement. In: Yegnanarayana B (ed) Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018. ISCA, pp 3244–3248, https://doi.org/10.21437/INTERSPEECH.2018-1400
2. Alcázar JL, Caba F, Mai L, et al (2020) Active speakers in context. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. Computer Vision Foundation / IEEE, pp 12462–12471, https://doi.org/10.1109/CVPR42600.2020.01248
3. Alcázar JL, Heilbron FC, Thabet AK, et al (2021) MAAS: multi-modal assignation for active speaker detection. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. IEEE, pp 265–274, https://doi.org/10.1109/ICCV48922.2021.00033
4. Alcázar JL, Cordes M, Zhao C, et al (2022) End-to-end active speaker detection. In: Avidan S, Brostow GJ, Cissé M, et al (eds) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVII, Lecture Notes in Computer Science, vol 13697. Springer, pp 126–143, https://doi.org/10.1007/978-3-031-19836-6_8
5. Carneiro HCC, Weber C, Wermter S (2021) FaVoA: Face-voice association favours ambiguous speaker detection. In: Farkaš I, Masulli P, Otte S, et al (eds) Artificial Neural Networks and Machine Learning - ICANN 2021 - 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14-17, 2021, Proceedings, Part I, Lecture Notes in Computer Science, vol 12891. Springer, pp 439–450, https://doi.org/10.1007/978-3-030-86362-3_36
6. Carneiro HCC, Weber C, Wermter S (2023) Whose emotion matters? Speaking activity localisation without prior knowledge. Neurocomputing 545:126271. https://doi.org/10.1016/J.NEUCOM.2023.126271
7. Chakravarty P, Mirzaei S, Tuytelaars T, et al (2015) Who's speaking?: Audio-supervised classification of active speakers in video. In: Zhang Z, Cohen P, Bohus D, et al (eds) Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 09 - 13, 2015. ACM, pp 87–90, https://doi.org/10.1145/2818346.2820780
8. Cho K, van Merrienboer B, Gülçehre Ç, et al (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Moschitti A, Pang B, Daelemans W (eds) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. ACL, pp 1724–1734, https://doi.org/10.3115/V1/D14-1179
9. Chung JS (2019) Naver at ActivityNet challenge 2019 - Task B active speaker detection (AVA). https://static.googleusercontent.com/media/research.google.com/en//ava/2019/Naver_Corporation.pdf
10. Chung JS, Lee B, Han I (2019) Who said that?: Audio-visual speaker diarisation of real-world meetings. In: Kubin G, Kacic Z (eds) Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019. ISCA, pp 371–375, https://doi.org/10.21437/INTERSPEECH.2019-3116
11. Chung JS, Huh J, Nagrani A, et al (2020) Spot the conversation: Speaker diarisation in the wild. In: Meng H, Xu B, Zheng TF (eds) Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020. ISCA, pp 299–303, https://doi.org/10.21437/INTERSPEECH.2020-2337
12. Cutler R, Davis LS (2000) Look who's talking: Speaker detection using video and audio correlation. In: 2000 IEEE International Conference on Multimedia and Expo, ICME 2000, New York, NY, USA, July 30 - August 2, 2000. IEEE Computer Society, pp 1589–1592, https://doi.org/10.1109/ICME.2000.871073
13. Datta G, Etchart T, Yadav V, et al (2022) ASD-Transformer: Efficient active speaker detection using self and multimodal transformers. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022. IEEE, pp 4568–4572, https://doi.org/10.1109/ICASSP43922.2022.9746991

14. Donley J, Tourbabin V, Lee J, et al (2021) EasyCom: An augmented reality dataset to support algorithms for easy communication in noisy environments. https://doi.org/10.48550/ARXIV.2107.04174
15. Grauman K, Westbury A, Byrne E, et al (2022) Ego4D: Around the world in 3,000 hours of egocentric video. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. IEEE, pp 18973–18990, https://doi.org/10.1109/CVPR52688.2022.01842
16. Hegde SB, Zisserman A (2023) GestSync: Determining who is speaking without a talking head. In: 34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023. BMVA Press, pp 506–509, http://proceedings.bmvc2023.org/506/
17. Howard AG, Zhu M, Chen B, et al (2017) MobileNets: Efficient convolutional neural networks for mobile vision applications. https://doi.org/10.48550/ARXIV.1704.04861
18. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Bach FR, Blei DM (eds) Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, JMLR Workshop and Conference Proceedings, vol 37. JMLR.org, pp 448–456, http://proceedings.mlr.press/v37/ioffe15.html
19. Jiang Y, Tao R, Pan Z, et al (2023) Target active speaker detection with audio-visual cue. In: Harte N, Carson-Berndsen J, Jones G (eds) Interspeech 2023, 24th Annual Conference of the International Speech Communication Association, Dublin, Ireland, 20 - 24 August 2023. ISCA, pp 3152–3156, https://doi.org/10.21437/INTERSPEECH.2023-574
20. Jung C, Lee S, Nam K, et al (2024) TalkNCE: Improving active speaker detection with talk-aware contrastive learning. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 8391–8395, https://doi.org/10.1109/ICASSP48485.2024.10448124
21. Kim YJ, Heo H, Choe S, et al (2021) Look who's talking: Active speaker detection in the wild. In: Hermansky H, Cernocký H, Burget L, et al (eds) Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021. ISCA, pp 3675–3679, https://doi.org/10.21437/INTERSPEECH.2021-2041
22. Köpüklü O, Taseska M, Rigoll G (2021) 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, Montreal, QC, Canada, October 10-17, 2021. IEEE, pp 1173–1183, https://doi.org/10.1109/ICCV48922.2021.00123
23. Liao J, Duan H, Feng K, et al (2023) A light weight model for active speaker detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. IEEE, pp 22932–22941, https://doi.org/10.1109/CVPR52729.2023.02196
24. Lin T, Maire M, Belongie SJ, et al (2014) Microsoft COCO: Common objects in context. In: Fleet DJ, Pajdla T, Schiele B, et al (eds) Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V, Lecture Notes in Computer Science, vol 8693. Springer, pp 740–755, https://doi.org/10.1007/978-3-319-10602-1_48
25. Min K, Roy S, Tripathi S, et al (2022) Learning long-term spatial-temporal graphs for active speaker detection. In: Avidan S, Brostow GJ, Cissé M, et al (eds) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXV, Lecture Notes in Computer Science, vol 13695. Springer, pp 371–387, https://doi.org/10.1007/978-3-031-19833-5_22
26. MMPose Contributors (2020) OpenMMLab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose
27. Qian X, Madhavi MC, Pan Z, et al (2021) Multi-target DoA estimation with an audio-visual fusion mechanism. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021. IEEE, pp 4280–4284, https://doi.org/10.1109/ICASSP39728.2021.9413776
28. Qian X, Brutti A, Lanz O et al (2022) Audio-visual tracking of concurrent speakers. IEEE Transactions on Multimedia 24:942–954. https://doi.org/10.1109/TMM.2021.3061800
29. Qu L, Weber C, Wermter S (2020) Multimodal target speech separation with voice and face references. In: Meng H, Xu B, Zheng TF (eds) Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020. ISCA, pp 1416–1420, https://doi.org/10.21437/INTERSPEECH.2020-1697
30. Radman A, Laaksonen J (2024) AS-Net: Active speaker detection using deep audio-visual attention. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-024-18457-9
31. Roth J, Chaudhuri S, Klejch O, et al (2020) AVA Active Speaker: An audio-visual dataset for active speaker detection. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020. IEEE, pp 4492–4496, https://doi.org/10.1109/ICASSP40776.2020.9053900
32. Roxo T, Costa JC, Inácio PRM et al (2024) WASD: A wilder active speaker detection dataset. IEEE Transactions on Biometrics, Behavior, and Identity Science. https://doi.org/10.1109/TBIOM.2024.3412821
33. Shahid M, Beyan C, Murino V (2021) S-VVAD: Visual voice activity detection by motion segmentation. In: IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021. IEEE, pp 2331–2340, https://doi.org/10.1109/WACV48630.2021.00238
34. Sharma R, Somandepalli K, Narayanan S (2023) Cross modal video representations for weakly supervised active speaker localization. IEEE Transactions on Multimedia 25:7825–7836. https://doi.org/10.1109/TMM.2022.3229975
35. Stefanov K, Sugimoto A, Beskow J (2016) Look who's talking: visual identification of the active speaker in multi-party human-robot interaction. In: Truong KP, Heylen D, Nishida T, et al (eds) Proceedings of the 2nd Workshop on Advancements in Social Signal Processing for Multimodal Interaction, ASSP4MI@ICMI 2016, Tokyo, Japan, November 12 - 16, 2016. ACM, pp 22–27, https://doi.org/10.1145/3005467.3005470
36. Stefanov K, Beskow J, Salvi G (2017) Vision-based active speaker detection in multiparty interaction. In: Salvi G, Dupont S (eds) Proceedings of GLU 2017 International Workshop on Grounding Language Understanding. ISCA, pp 47–51, https://doi.org/10.21437/GLU.2017-10
37. Sun K, Xiao B, Liu D, et al (2019) Deep high-resolution representation learning for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, pp 5693–5703, https://doi.org/10.1109/CVPR.2019.00584
38. Tao R, Pan Z, Das RK, et al (2021) Is someone speaking?: Exploring long-term temporal features for audio-visual active speaker detection. In: Shen HT, Zhuang Y, Smith JR, et al (eds) MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021. ACM, pp 3927–3935, https://doi.org/10.1145/3474085.3475587

39. Vajaria H, Sarkar S, Kasturi R (2008) Exploring co-occurence between speech and body movement for audio-guided video localization. IEEE Transactions on Circuits and Systems for Video Technology 18(11):1608–1617. https://doi.org/10.1109/TCSVT.2008.2005602
40. Wang X, Cheng F, Bertasius G, et al (2024) LoCoNet: Long-short context network for active speaker detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 17-21, 2024. Computer Vision Foundation / IEEE, pp 18462–18472, https://openaccess.thecvf.com/content/CVPR2024/html/Wang_LoCoNet_Long-Short_Context_Network_for_Active_Speaker_Detection_CVPR_2024_paper.html
41. Wuerkaixi A, Zhang Y, Duan Z, et al (2022) Rethinking audio-visual synchronization for active speaker detection. In: 32nd IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2022, Xi'an, China, August 22-25, 2022. IEEE, pp 1–6, https://doi.org/10.1109/MLSP55214.2022.9943352
42. Xiong J, Zhou Y, Zhang P et al (2023) Look &listen: Multi-modal correlation learning for active speaker detection and speech enhancement. IEEE Transactions on Multimedia 25:5800–5812. https://doi.org/10.1109/TMM.2022.3199109
43. Yan S, Xiong Y, Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. AAAI Press, pp 7444–7452, https://doi.org/10.1609/AAAI.V32I1.12328
44. Zhang Y, Xiao J, Yang S, et al (2019) Multi-task learning for audio-visual active speaker detection. https://static.googleusercontent.com/media/research.google.com/en//ava/2019/Multi_Task_Learning_for_Audio_Visual_Active_Speaker_Detection.pdf
45. Zhang Y, Liang S, Yang S, et al (2021a) ICTCAS-UCAS-TAL submission to the AVA-ActiveSpeaker task at ActivityNet Challenge 2021. https://static.googleusercontent.com/media/research.google.com/en//ava/2021/S1_ICTCAS-UCAS-TAL.pdf
46. Zhang Y, Liang S, Yang S, et al (2021b) UniCon: Unified context network for robust active speaker detection. In: Shen HT, Zhuang Y, Smith JR, et al (eds) MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021. ACM, pp 3964–3972, https://doi.org/10.1145/3474085.3475275
47. Zhang Y, Liang S, Yang S, et al (2022) UniCon+: ICTCAS-UCAS submission to the AVA-ActiveSpeaker task at ActivityNet Challenge 2022. https://static.googleusercontent.com/media/research.google.com/en//ava/2022/S1_ICTCAS_UCAS_UniCon+.pdf

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.