Radman, Abduljalil; Laaksonen, Jorma

AS-Net : active speaker detection using deep audio-visual attention

Published in:
Multimedia Tools and Applications

DOI:
10.1007/s11042-024-18457-9

Published: 01/08/2024

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Radman, A., & Laaksonen, J. (2024). AS-Net : active speaker detection using deep audio-visual attention. *Multimedia Tools and Applications*, *83*(28), 72027-72042. https://doi.org/10.1007/s11042-024-18457-9

Multimedia Tools and Applications (2024) 83:72027–72042
https://doi.org/10.1007/s11042-024-18457-9

AS-Net: active speaker detection using deep audio-visual attention

Abduljalil Radman[1] · **Jorma Laaksonen**[1]

Received: 30 May 2023 / Revised: 26 January 2024 / Accepted: 27 January 2024 /
Published online: 5 February 2024

Abstract

Active Speaker Detection (ASD) aims at identifying the active speaker among multiple speakers in a video scene. Previous ASD models often seek audio and visual features from long video clips with a complex 3D Convolutional Neural Network (CNN) architecture. However, models based on 3D CNNs can generate discriminative spatial-temporal features, but this comes at the expense of computational complexity, and they frequently face challenges in detecting active speakers in short video clips. This work proposes the **A**ctive **S**peaker **Net**work (AS-Net) model, a simple yet effective ASD method tailored for detecting active speakers in relatively short video clips without relying on 3D CNNs. Instead, it incorporates the Temporal Shift Module (TSM) into 2D CNNs, facilitating the extraction of dense temporal visual features without the need for additional computations. Moreover, self-attention and cross-attention schemes are introduced to enhance long-term temporal audio-visual synchronization, thereby improving ASD performance. Experimental results demonstrate that AS-Net outperforms state-of-the-art 2D CNN-based methods on the AVA-ActiveSpeaker dataset and remains competitive with the methods utilizing more complex architectures.

Keywords Active speaker detection · Audio-visual attention · Temporal shift module · Audio-visual features · Convolutional Neural Networks (CNNs)

1 Introduction

Understanding video scenes is fundamental in numerous multimedia and computer vision applications. Active Speaker Detection (ASD) has emerged as a promising solution within video scene understanding, focusing on the identification of an active speaker among potential speakers in a visual scene. This task is particularly challenging due to the complex interplay of audio and visual information in video scenes. Recently, ASD research has received

✉ Abduljalil Radman
abduljalil.saif@aalto.fi

Jorma Laaksonen
jorma.laaksonen@aalto.fi

[1] Department of Computer Science, Aalto University, Espoo, Finland

significant attention in many disciplines, including biometrics [1], speaker tracking [2, 3], speaker diarization [4–6], and speech understanding [7, 8]. Despite of several promising ASD approaches in the literature, it remains an unsolved research topic.

Earlier ASD studies faced limitations, either relying on audio or visual unimodal models [9–12] or utilizing constrained datasets [13]. While ASD methods based on unimodal audio models provided valuable insights for speaker diarization [10, 14], unimodal visual models (primarily tracing face movements) demonstrated strong performance in single-speaker scenarios [12]. However, audio unimodal models face challenges from background noise and off-screen speakers. Visual unimodal models, are susceptible to detection errors due to insufficient features from the face region, such as low resolution and occlusion, or interference from non-speaking activities like yawning or grinning [15]. On top of this, the majority of conventional ASD approaches have been developed using data from controlled laboratory settings [14], making it challenging to generalize these methods effectively for real-world applications.

With the release of the AVA-ActiveSpeaker dataset [16], that is an audio-visual large dataset with diversity in demographics, illuminations, face occlusions, recording device resolutions, sampling rates, on and off-screen speakers, and face sizes, it has become possible to develop ASD for real-world scenarios. Over and above that, neurobiological studies have highlighted numerous perceptual advantages of visual-auditory interactions on improving predominantly unimodal brain regions in multi-sensory processing [17]. Thus, the current trend of ASD research is to leverage integration between the audio and visual information to implement robust multi-modal ASD models. This involves formulating the ASD problem as a multi-modal model that exploits the synchronization of subtle facial movements with corresponding audio snippets.

Visual cues, as highlighted in existing literature [3, 18], provide complementary information to audio cues. Integrating them into a multi-modal ASD model addresses two key challenges: 1) the multi-speaker detection problem that requires careful face-voice synchronization to avoid false detection of facial movements which resembling speaking activities (e.g., laughing and eating), and 2) the temporal consistency between the audio and visual information, which rapidly evolves over time (e.g., a dialogue among three speakers) [15, 19].

Most of the audio-visual ASD approaches in the literature address the ASD problem by first extracting the visual (i.e., mainly facial) features and the audio features from short video clips, and then learning the long-term temporal synchronization based on these audio-visual features. The audio and visual features are commonly extracted via 2D Convolutional Neural Networks (CNNs) [15, 16, 19–24] or 3D CNNs [25–29], whereas Recurrent Neural Networks (RNNs) [30] are used to provide the temporal synchronization of the audio-visual features. While 2D CNNs excel at capturing precise spatial features, they fall short in inferring temporal relationships. On the other hand, 3D CNNs can simultaneously learn spatio-temporal features, but their approach is computationally intensive and demands substantial memory resources [24, 31]. The majority of existing ASD models follow a common practice of extracting audio and visual features from short video clips, predominantly utilizing 2D CNNs. The primary distinction among these models lies in their approaches for modeling long-term audio-visual temporal synchronization. RNNs, such as LSTM [32] or GRU [33], are usually applied following 2D CNNs to model the audio-visual temporal synchronization [16, 20, 25]. However, RNNs may fail to capture the low-level information that is lost by the 2D CNN feature extractors. As an alternative way to handle the long-term temporal information, attention-based mechanisms [27, 28] offer good performance. Attention mechanisms [34] are crucial in machine learning models, selectively emphasizing dominant features and refining discriminative ones to enhance salient aspects of input data. They enable prioritization and assign more weight to influential elements [35], guiding processes like action recognition for pertinent temporal segments [36] and refining text generation in image and video captioning by accentuating essential visual elements [37]. Graph Neural Networks (GNNs) have also been used for modeling long-term temporal audio-visual synchronization [19, 23, 24]. Although GNN-based approaches prove effective in achieving long-term temporal synchronization, their architectural complexity tends to be more intricate.

In this paper, a simple yet effective ASD method, depicted in Fig. 1, has been proposed and given the name **A**ctive **S**peaker **Net**work (AS-Net). AS-Net is built upon the AVA-ActiveSpeaker baseline architecture with two-stream modality fusion [16] for extracting audio and visual features. It specifically focuses on enhancing temporal feature extraction from relatively short video clips, employing 2D CNNs instead of more complex 3D CNNs. To enhance the temporal feature extraction capacity of 2D CNNs, akin to 3D CNNs, AS-Net incorporates the light-weight Temporal Shift Module (TSM) [31], facilitating improved temporal feature extraction without introducing additional computational complexity. Next, AS-Net focuses on the long-term audio-visual temporal information, employing both cross and self-attention schemes to effectively address synchronization across the temporal dimension. Finally, AS-Net predicts the presence of active speakers in a video scene given this long-term temporal audio-visual information. The contributions of AS-Net are threefold: (1) Demonstrating the viability and preference of learning dense visual features from short video clips by incorporating TSM into 2D CNNs. (2) Enhancing long-term temporal audio-visual synchronization through the implementation of simple cross and self-attention schemes. (3) Notably outperforming state-of-the-art 2D CNN-based models and maintaining competitiveness with models built upon deeper architectures.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 explains in detail the proposed AS-Net method. Experimental results are presented and analyzed in Section 4, while Section 5 concludes the work.

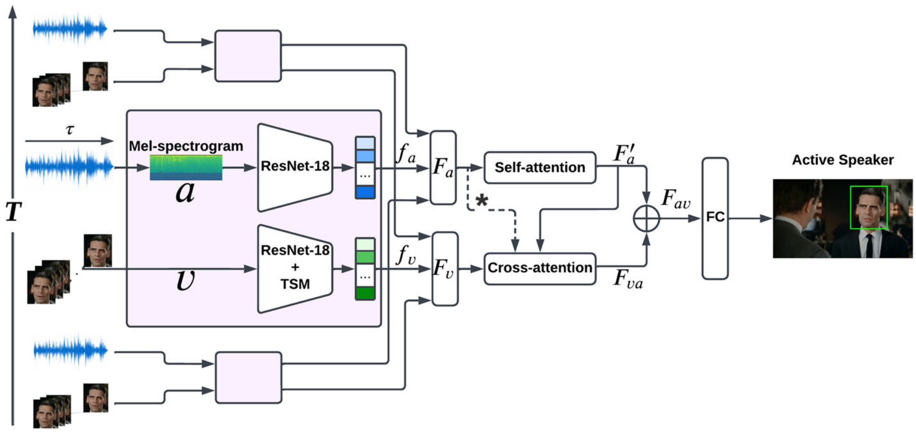


Fig. 1 The proposed AS-Net begins by extracting the visual $\boldsymbol{f_v}$ and audio $\boldsymbol{f_a}$ features for the target speaker from a stack of face crops $\boldsymbol{v}$ and their corresponding audio snippet $\boldsymbol{a}$ in short video clips of duration τ, receptively. Then, it aggregates these visual and audio features over a longer window T ($T \gg \tau$) to model the long-term temporal audio-visual synchronization $\boldsymbol{F_{av}}$ using self and cross-attention schemes. Finally, the $\boldsymbol{F_{av}}$ representation is used to discriminate the active speaker from non-active speakers. The symbol * denotes an alternative approach for performing audio-visual cross-attention

2 Related work

This section explores how closely-related works handle the extraction of audio and visual features and address long-term temporal audio-visual synchronization in the ASD task. Early works of the ASD problem [9, 38] rely on analyzing the audio information only. However, the background noise, such as overlapping speech, clapping, and music, may lead to false ASD. On the other hand, visual-based ASD approaches [11, 12], which hinge on face movement detection, can lead to detection errors owing to insufficient features from the face region (e.g., low resolution and occlusion) or the presence of non-speaking activities such as yawning, licking lips, or grinning. In addition, some words are pronounced solely through tongue movements, which do not induce corresponding face movements [39]. Importantly, the visual cues carry complementary information to the audio cues, and the integration between them can tackle the problems of audio and visual unimodal models as well as potentially increase the ASD performance [15, 16, 39]. Hence, recent approaches fuse the audio and visual information for reliable ASD performance.

Zhang et al [40] employed a large 3D CNN to extract per-frame audio and visual features. These features were then concatenated and fed into a two-layer temporal GRU model to derive final frame-wise predictions. Visual features were extracted using the 3D-ResNet-18 network, focusing on the lips' visual characteristics, while audio features were obtained from 13-dimensional Mel-frequency cepstral coefficients (MFCCs) through SyncNet [41]. In [20], the visual features from five consecutive face frame inputs were encoded using the 3D VGG-M [42] network, while the corresponding audio inputs (i.e., MFCCs) were encoded using a 2D CNN encoder. Subsequently, the audio and visual features were input into distinct Bi-LSTM [43] networks. The outputs from these networks were concatenated through a linear classification layer to predict the presence of an active speaker. TalkNet [27] also relied on a 3D encoder for visual representation and a 2D encoder for audio representation, and analyzed the scene's contextual information by means of an attention technique [34] instead of RNNs. In contrast to previous works, relatively long video clips (~ 2 s) were used for audio and visual feature extraction with the claim that is hard to distinguish speaking activity from non-speaking activity using short video clips ($\sim 200 - 600$ ms). ASDNet [25] employed a 3D CNN encoder as a visual encoder and a 2D CNN encoder as an audio encoder. ASDNet extracted the audio features directly from the raw audio signals rather than from MFCCs. Moreover, information from background speakers was aggregated into the target speaker information to enhance the audio-visual representation. Several RNN scenarios were explored to study the coherency of speaking activity in the ASDNet work. In [44], a short context network was designed to capture both long-term intra-speaker context and short-term inter-speaker context. The model utilized self-attention for the effective representation of long-range dependencies within intra-speaker context and incorporated 3D convolutional blocks to capture local patterns, addressing short-term inter-speaker context. While 3D CNN-based methods excel in extracting robust spatial-temporal features from a collection of facial thumbnails, they come with the drawbacks of complexity, high computational demands, and substantial memory usage [24].

In the same context, GNNs were introduced to jointly model the long-term audio-visual temporal synchronization for ASD [19, 23, 40]. Alcázar et al. [19] proposed the Multi-modal Assignation for Active Speaker (MAAS) detection model that relies on multi-modal GNNs. MAAS used a short temporal window (~ 1.6 s) to construct a GNN for handling the temporal context of audio-visual synchronization. In this method, speakers were connected on the graph only if they appeared in consecutive frames. On the contrary, [24] relaxed the connectivity among speakers to a long window spanning from 13 to $\sim$ 55 s. They also applied various temporal ordering patterns (forward, backward, undirected) to enhance audio-visual synchronization and inter-speaker connections in the graph. The End-to-end Active Speaker dEtEction (EASEE) model, proposed by [23], learns multi-modal features from multiple visual tracklets and models their spatio-temporal relations in an end-to-end manner. The authors incorporated an interleaved GNN block to capture the relationship among speakers in consecutive frames, aiming for robust spatial-temporal synchronization. Although GNN-based models learn better inter-speaker and audio-visual temporal synchronization, the RNNs and attention-based techniques and their variants maintain a simpler architectural design.

Due to their simplicity and low memory requirements, 2D CNN-based models have also been widely used for the ASD task. In the work by [16], facial and audio features were learned through the optimization of a 2D CNN multi-modal encoder. This encoder served as a feature extractor for the subsequent stage, where embeddings from multiple speakers were fused and classified. Active Speaker Context (ASC) [15] used the 2D ResNet-18 [45] model to extract audio and visual features, and then applied a non-local attention module with LSTM to model the long-term synchronization between audio and visual features. Instead of holistically encoding the long-term multi-speaker relationship, a bottom-up scenario was proposed to learn this relationship through aggregating the audio and visual short-term features and mapping them into an embedding that explicitly models the pairwise relationship with the long-term audio-visual synchronization. Carneiro et al. [21] incorporated Face-Voice Association (FaVoA) [46] into the ASC model to address speaking activity detection failures arising from low speaker mouth resolution or the absence of the speaker's face. In another work by [22], a Unified Context (UniCon) network was proposed for the ASD task. UniCon jointly modeled various types of contextual information, including spatial, relational, and temporal contexts. This approach aimed to capture the scale and position of each speaker's face, synchronize candidate speakers, and alleviate synchronization errors between audio and visual streams. UniCon was implemented as a two-stream network, where a 2D ResNet-18 visual stream generates an average-pooled visual feature vector from consecutive face frames, and a 2D ResNet-18 audio stream produces a feature vector from MFCCs that exactly matches the time interval of the face frames. In general, models based on 2D CNNs offer lower computational costs and are well-suited for real-time applications.

Our simple and effective AS-Net method is also based on the 2D CNN architecture, particularly ResNet-18 [45], and focuses on boosting the short-term visual features and the long-term audio-visual synchronization. Unlike previous methods [24, 27, 29], which rely on audio augmentation with negative samples to improve the short-term audio-visual representation, our AS-Net employs the Temporal Shift Module (TSM) [31] to extract dense temporal visual features with reduced supervision. Furthermore, it learns the long-term audio-visual synchronization by self-attention and cross-attention mechanisms for improved ASD.

3 AS-Net

Our AS-Net, as outlined in Fig. 1, detects active speakers in a video by extracting audio and visual features from short video clips. It then models long-term temporal audio-visual synchronization using attention schemes.

3.1 Audio and visual feature extraction

Drawing inspiration from existing ASD approaches, we introduce a two-stream convolutional encoder that focuses on both audio and visual components to extract spatial-temporal features from short video clips of duration τ seconds. The audio stream input is a Mel-spectrogram $\boldsymbol{a} \in \mathbb{R}^{x \times y \times 1}$ calculated from the audio signal over τ, where x and y depend on τ. Meanwhile, the visual stream input $\boldsymbol{v} \in \mathbb{R}^{k \times 3 \times h \times w}$ is represented with k RGB face crops associated with the audio signal, where h and w are the height and width of k face crops. The k face crops are temporally sampled over the τ interval. The two-stream encoder encodes $\boldsymbol{a}$ and $\boldsymbol{v}$ into low-dimensional audio $\boldsymbol{f_a} \in \mathbb{R}^{d_a}$ and visual $\boldsymbol{f_v} \in \mathbb{R}^{d_v}$ embeddings, respectively.

The audio stream encoder uses a 2D ResNet-18 [45] to encode 2D Mel-spectrogram inputs through averaging channel weights at the input layer, while the visual stream encoder uses a 2D ResNet-18 incorporated with TSM [31] to efficiently acquire the temporal relationship among consecutive face crops. TSM is a plug-and-play module capable of handling complex temporal relationships without additional computation [31]. In this configuration, the TSM model is incorporated into each residual block of ResNet-18 before the first convolutional layer as shown in Fig. 2a. Following the approach in the original work [31], bi-directional TSM (Fig. 2b) with a small proportion of residual shift is utilized. The PyTorch code in Algorithm 1 provides a straightforward implementation of TSM.

Algorithm 1 Temporal Shift Module (TSM)

```
Input: $\mathbf{X} \in \mathbb{R}^{B \times D \times C \times H \times W}$, where $B$ is the batch size, $D$ is the temporal dimension, $C$ is the number of channels, $H$ and $W$ are the spatial resolutions of the channels, and $fold\_div$: proportion of the residual block channels.
$\mathbf{Y}$= torch.zeros_like($\mathbf{X}$).
$fold = C \,//\, fold\_div$.
$\mathbf{Y}[:, : -1, : fold] = \mathbf{X}[:, 1 :, : fold]$
$\mathbf{Y}[:, 1 :, fold : 2 * fold] = \mathbf{X}[:, : -1, fold : 2 * fold]$
$\mathbf{Y}[:, :, 2 * fold :] = \mathbf{X}[:, :, 2 * fold :]$
Output: $\mathbf{Y}$
```

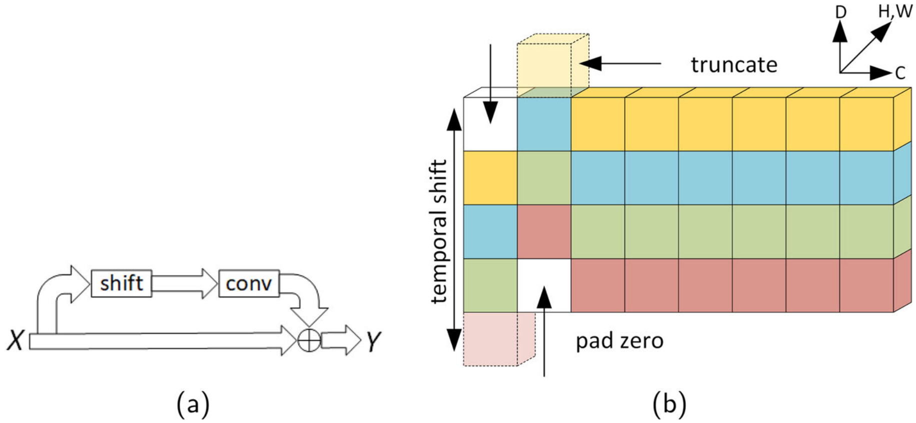


Fig. 2 (a) Residual Temporal Shift Module (TSM). (b) Bi-directional TSM

3.2 Long-term temporal audio-visual synchronization

The coherence of speaking activities in terms of time is essential, ensuring that the speaker maintains consistency across different time instants. This means that the person speaking in previous or future moments is likely the same speaker as at the current time instant. Audio-visual synchronization plays a crucial role in this regard. Hence, AS-Net aggregates low-dimensional audio $\boldsymbol{f_a}$ and visual $\boldsymbol{f_v}$ embeddings over a long temporal window T to yield $\boldsymbol{F_a} \in \mathbb{R}^{L\times d_a}$ and $\boldsymbol{F_v} \in \mathbb{R}^{L\times d_v}$ embeddings, respectively, where L is the number of short video clips uniformly sampled across the window T. $\boldsymbol{F_a}$ and $\boldsymbol{F_v}$ are then used to model the long-term temporal audio-visual synchronization by means of self and cross-attention mechanisms [34].

Audio self-attention. The dominant audio snippets for ASD are generally distributed over a long time interval of duration T. Thereby, AS-Net adopts a self-attention network featuring a multi-head attention layer (Fig. 3a) to capture these dominant snippets. The input for the attention layer, including the query Q_a, key K_a, and value V_a, is derived from the trainable linear embedding $\boldsymbol{F_a}$. The output, denoted as $\boldsymbol{F'_a}$ (representing the dominant audio snippets), is calculated as [34]:

$$\boldsymbol{F'_a} = softmax(\frac{Q_a K_a^T}{\sqrt{d_a}})V_a. \tag{1}$$

Audio-visual cross-attention. False positive speaking activities (e.g., yawning and laughing) may lead to false ASD. To alleviate this issue, AS-Net proposes a cross-attention network (Fig. 3b) to tightly link the visual cues in $\boldsymbol{F_v}$ with the predominant audio snippets $\boldsymbol{F'_a}$ across the temporal dimension T. The input query Q'_a for this cross-attention network is derived from the dominant audio snippets $\boldsymbol{F'_a}$, while the inputs for key K_v and value V_v are obtained

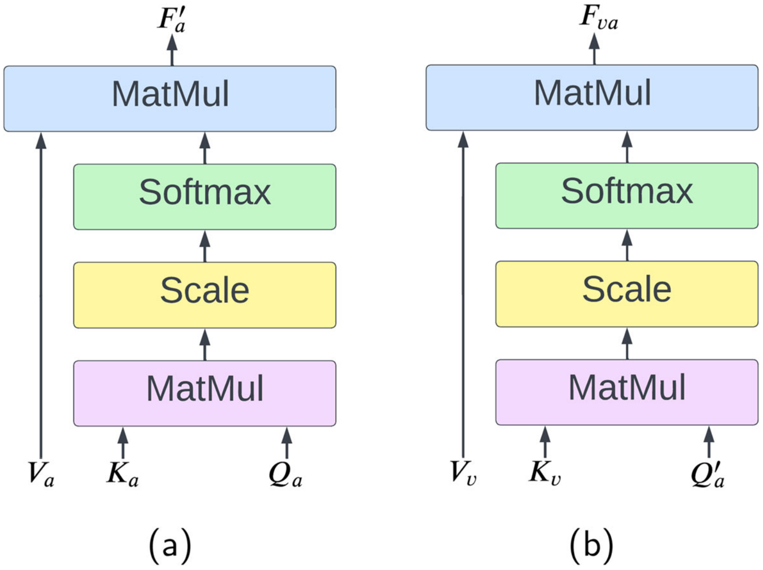


Fig. 3 (a) Audio self-attention. (b) Audio-visual cross-attention

from the visual embedding $\boldsymbol{F_v}$. The resulting output of the cross-attention network is the visual attention embedding $\boldsymbol{F_{va}}$ (Fig. 1) calculated as [34]:

$$\boldsymbol{F_{va}} = softmax(\frac{Q'_a K_v^T}{\sqrt{d_v}})V_v, \tag{2}$$

where $d_v = d_a$ is the dimensionality of the embeddings.

Finally, AS-Net concatenates the features $\boldsymbol{F_{va}}$ and $\boldsymbol{F'_a}$, creating a unified embedding $\boldsymbol{F_{av}} \in \mathbb{R}^{L\times(d_a+d_v)}$, and estimates the presence of active speaker given $\boldsymbol{F_{av}}$. Two fully connected layers are applied to project $\boldsymbol{F_{av}}$ onto the categories of active speaker and non-active speaker. The scores for the active speaker classes are determined using a Softmax operator.

3.3 Implementation details

Two-stream encoder. Following [15, 16], a two-stream convolutional encoder, specifically a 2D ResNet-18, was used to extract the audio and visual spatial-temporal embeddings from short video clips of duration τ= 407 ms. Before extracting audio and visual features, all short video clips underwent resampling to a frame rate of 27 fps. The audio stream encoder was designed to accept 2D Mel-spectrogram tensors calculated from the audio signal spanning τ. The audio signal underwent conversion into Mel-spectrogram bands in $\mathbb{R}^{x\times y\times 1=13\times 40\times 1}$, computed over a duration of 407 ms with a sampling rate of 16000, a window size of 25 ms, a hop length of 10 ms, and 13 coefficients. The visual stream encoder takes as input a set of $k = 11$ face crops, which exactly match the audio stream input. To enhance the encoding of temporal features from these k face crops (denoted as $D = 11$ in Algorithm 1), we integrated the TSM model [31] into the visual stream encoder. We adopted a 1/4 proportion bi-directional residual shift (1/8 for each direction) to effectively capture informative temporal features (i.e., $fold = 8$ in Algorithm 1 when the residual block contains 64 channels). The resulting embeddings from both the audio and visual streams are 512-dimensional ($d_a = d_v = 512$).

Two-stream encoder training. Each face crop (the visual stream inputs) was re-sized to $h\times w = 144\times 144$. Additionally, visual augmentation techniques, including flipping, scaling, and color jittering, were applied to each face crop. On the contrary, the raw data of Mel-spectrograms (the audio stream inputs) were employed without any additional augmentation. Training the two-stream encoder involved the use of the ADAM optimizer [47] with a learning rate of 3×10^{-4}, learning rate annealing (0.1 for every 40 epochs), and a batch size of 64, spanning a training period of 100 epochs. The loss function $\mathcal{L}$ to obtain the audio and visual embeddings was calculated from three cross entropy losses as follows:

$$\mathcal{L} = \mathcal{L}_a + \mathcal{L}_v + \mathcal{L}_{av}, \tag{3}$$

where $\mathcal{L}_a$ and $\mathcal{L}_v$ are the cross entropy losses based on the audio and visual streams predictions, respectively, and $\mathcal{L}_{av}$ is the cross entropy based on the prediction of audio-visual embedding (i.e., formed by fusing the audio and visual embeddings).

Self and cross-attention networks. Both the self-attention and cross-attention networks were constructed from one attention layer with eight attention heads for modeling the long-term temporal audio-visual synchronization. The inputs to these networks were sampled from the audio and visual embeddings (i.e., obtained from the two-stream encoder) over a window of size T. Fifteen video clips $L = 15$ were adopted to yield a tensor $\mathbb{R}^{15\times 512}$ (i.e., $T = 3$ s for 20 fps videos and $T = 2$ s for 30 fps videos). For training the model over 25 epochs, an ADAM optimizer with an initial learning rate of 3×10^{-6}, a learning rate annealing of

0.1 every 5 epochs, and a batch size of 64 were utilized. A single cross entropy loss function based on $\boldsymbol{F_{av}}$ in $\mathbb{R}^{15\times1024}$ embedding was adopted to train the final model.

4 Experimental results and analysis

In this section, the proposed AS-Net method is evaluated with the AVA-ActiveSpeaker dataset [16]. First, the large-scale AVA-ActiveSpeaker dataset is presented and analyzed. AS-Net is then compared with the state-of-the-art ASD approaches. Next, ablation studies are conducted to show the contribution of each part of the AS-Net method. Finally, qualitative analysis is presented to show the robustness of the proposed AS-Net with visual and audio noise.

4.1 AVA-ActiveSpeaker dataset

The AVA-ActiveSpeaker dataset [16] is the first large and diverse active speaker dataset collected in the wild. It is an audio-visual dataset derived from 297 Hollywood movies, 133 for training, 33 for validation, and 131 for testing. It comprises of about 3.65 million human labeled frames spanning 38.5 hours of face tracks and the corresponding audio snippets. Each face track in the AVA-ActiveSpeaker dataset is accompanied by augmented attributes indicating whether the person is *speaking* or *non-speaking*. This dataset presents challenges due to its diversity in language, face resolution, frame rate, and audible background noise [16, 27].

4.2 Comparison to state-of-the-art

For fair comparison with state-of-the-art methods, AS-Net is compared against 2D CNN based methods, given its reliance on the 2D CNN architecture. The performance of AS-Net is compared with the published performances of AVA [16], ASC [15], FaVoA [21], MAAS [19], UniCon (UniCon+T, UniCon+S+T, UniCon+R+T and UniCon+S+R+T, where T: temporal context, S: spatial context, and R: relational context) [22], and EASEE-2D [23]. All method performances are assessed on the AVA-ActiveSpeaker validation subset using the official evaluation tool provided with the dataset. The mean average precision (mAP) metric is employed for this comparative analysis.

Table 1 illustrates the comparative performance of various models on the AVA-Active Speaker validation subset. It shows that FaVoA [21] gives the worst performance on the AVA-ActiveSpeaker validation subset. It also shows that ASC [15] outperforms FaVoA, despite FaVoA being built upon the ASC model. As depicted in Table 1, MAAS [19] surpasses FaVoA and ASC models, demonstrating its superior capability in effectively handling long-term temporal audio-visual synchronization using GNNs. Likewise, EASEE-2D [23] leverages GNNs for modeling the temporal audio-visual synchronization. Thus, EASEE-2D overcomes all preceding methods including UniCon+T, UniCon+S+T, and UniCon+R+T [22], but it performs worse than UniCon+S+R+T [22]. UniCon+S+R+T employs a complex architecture involving multiple 2D CNN backbones and Bi-GRUs to analyze long-term temporal audio-visual synchronization. Notably, this model requires extra supervision data, such as spatial information about the speakers. In contrast, the proposed AS-Net surpasses the performance of EASEE-2D by 0.30 mAP with less supervision information and a simpler architecture (2D CNN and attention schemes without the use of GNNs or RNNs). Remarkably, AS-Net

Table 1 Comparison between the proposed AS-Net with state-of-the-art methods on the AVA-ActiveSpeaker validation subset [16], in terms of mAP

Method	Audio-visual encoder	Temporal synchronization	mAP
ASC [15]	ResNet-18	PW-A+LSTM	87.10
FaVoA [21]	ResNet-18	PW-A+LSTM+GRU	84.70
MAAS [19]	ResNet-18	GNN	88.80
UniCon+T [22]	ResNet-18+VGG-M	Multiple	89.60
UniCon+S+T [22]	ResNet-18+VGG-M	Multiple	90.30
UniCon+R+T [22]	ResNet-18+VGG-M	Multiple	90.40
UniCon+S+R+T [22]	ResNet-18+VGG-M	Multiple	92.00
EASEE-2D [23]	ResNet-18	GNN	91.10
AS-Net (Ours)	ResNet-18	Cross and self-attention	91.40

”PW-A” denotes Pairwise-attention

achieves a significant improvement of 1.8 mAP over the leading approach with a similar architecture, namely UniCon+T.

4.3 Ablation study

In this section, ablation studies are conducted to validate the significance of AS-Net's components. Table 2 summarizes the contribution of AS-Net's components in terms of mAP. The abbreviations used in Table 2 stand for Audio-Visual fusion before applying the attention schemes (AV), Audio Self-Attention (ASA), and Audio-Visual Cross-Attention (AVCA).

Impact of the TSM on performance. Table 2 provides a comparison of the two-stream encoder's performance both before and after incorporating the TSM model. Although the TSM model was exclusively incorporated on the visual stream encoder, the evaluation of the two-stream encoder's performance is based on audio-visual embedding prediction. This involves concatenating the outputs of both streams, denoted as f_a and f_v, and utilizing the loss function defined in (3). The results reveal a significant improvement in the two-stream

Table 2 Contribution of AS-Net's components to the final performance

TSM	AV	ASA	AVCA	mAP
✗	✗	✗	✗	79.61
✓	✗	✗	✗	86.32
✓	✓	✗	✗	89.69
✓	✓	✓	✗	89.91
✓	✓	✗	✓	90.96
✓	✓	✓	✓ *	91.31
✓	✓	✓	✓	91.40

AV indicates the absence of any attention scheme, ASA indicates the application of the Audio Self-Attention scheme, and AVCA indicates the application of the Audio-Visual Cross-Attention scheme

The symbol * indicates that the AVCA is the approach marked with * in Fig. 1

Here, $\tau = 407$ ms, $k = 11$, $T = 3$, and $L = 15$

Table 3 Impact of the clip length τ on the two-stream encoder performance

Clip length	$\sim$ 200 ms, $k = 5$	$\sim$ 400 ms, $k = 11$	$\sim$ 800 ms, $k = 21$
mAP	80.98	86.32	87.27

encoder's performance following the incorporation of the TSM model. Specifically, TSM contributes to a notable increase of 6.71 in mAP. This is attributed to the power of the TSM model in learning temporal features from k consecutive face crops, leveraging the simplicity of the 2D ResNet-18 architecture. Throughout subsequent ablation experiments, the TSM model is consistently paired with 2D ResNet-18 to compose the two-stream encoder.

Impact of the clip length τ on performance. Despite training the two-stream encoder with short video clips ($\tau = 407$ ms, $k = 11$), it is noteworthy that longer video clips may lead to enhanced performance, as evident in Table 3 (e.g., achieving 87.27 mAP with $k = 21$). However, we sacrificed this improvement to keep our model simple because long video clips require more memory budget and also hinder ASD in shorter video clips. For the rest of the ablation studies, we consistently utilize $k = 11$.

Impact of the window length T on performance. In this ablation study, we investigate the impact of the temporal window length T on generating embeddings $\boldsymbol{F_a}$ and $\boldsymbol{F_v}$. The mAP values (Table 4) are reported by fusing $\boldsymbol{F_a}$ and $\boldsymbol{F_v}$ over varying window lengths T before applying attention schemes (i.e., AV in Table 2). Table 4 highlights the crucial role of T in enhancing the performance of AS-Net. Notably, the window length of $T = 3$ s leads to a performance improvement of 3.37 mAP compared to the two-stream encoder performance (TSM in Table 2). Furthermore, it is evident that $T = 3$ s outperforms both $T = 1$ s and $T = 5$ s. Consequently, we adopt $T = 3$ s in our experiments, unless specified otherwise.

Impact of the number of short video clips L on performance. In all previous experiments, we consistently employed $L = 15$ short video clips within each window T. Decreasing L to 5 resulted in a lower performance, yielding an mAP of 89.18, whereas increasing L to 21 led to slower execution with a negligible improvement of 0.05 mAP as depicted in Table 5. Therefore, we maintain the use of $L = 15$ in our experiments.

Effect of the audio self-attention on performance. The contribution of the audio self-attention network to AS-Net's final performance is evident, as it results in a 0.22 mAP improvement (ASA in Table 2) compared to the AV performance in the same table. This provides evidence that the audio self-attention network is capable of capturing dominant audio snippets, facilitating long-term temporal audio-visual synchronization.

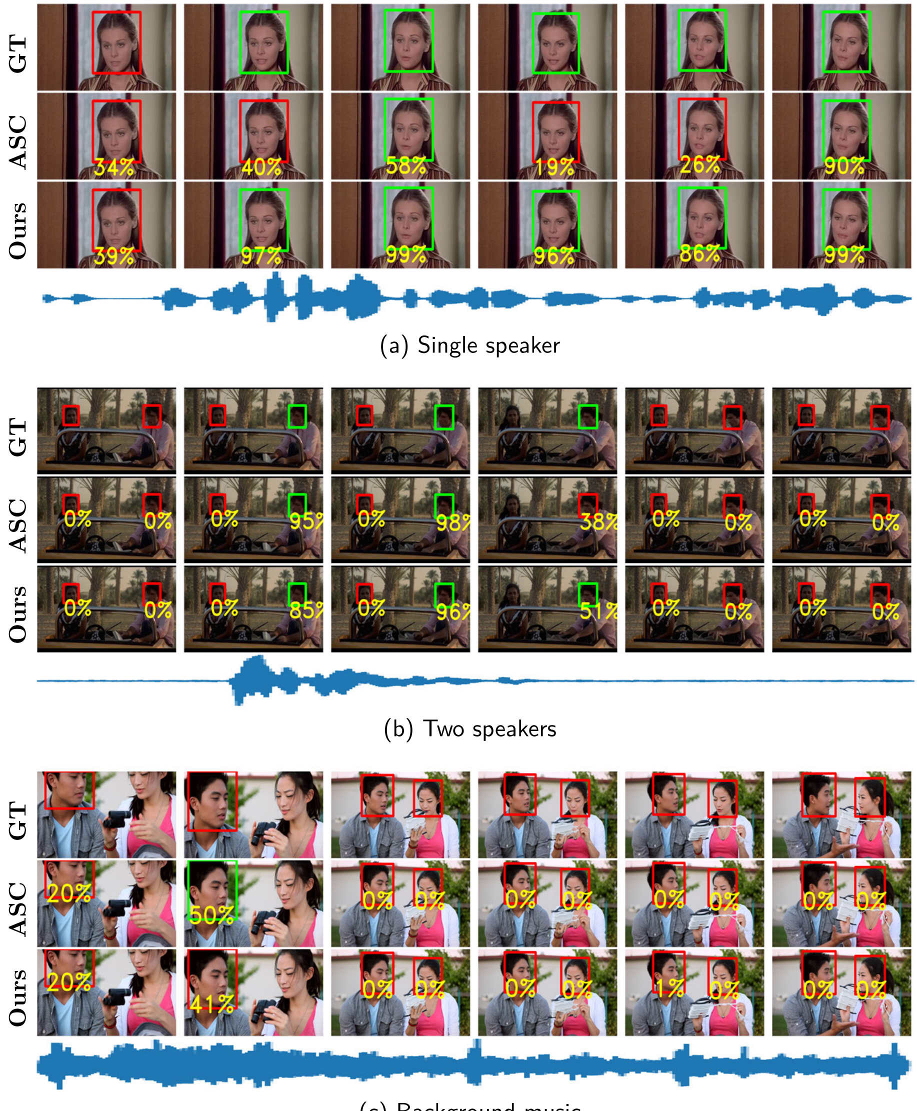
Effect of the audio-visual cross-attention on performance. The integration of the audio-visual cross-attention network significantly enhances AS-Net's performance, leading to a 1.27 mAP improvement (AVCA in Table 2). This outcome underscores the effectiveness of the cross-attention scheme in learning and leveraging the correlation between dominant audio snippets and corresponding facial movements for improved long-term temporal audio-visual synchronization.

Table 4 Impact of the window length T on performance

Window length	$T = 1$ s	$T = 3$ s	$T = 5$ s
mAP	89.07	89.69	89.66

Table 5 Impact of the number of short video clips L on performance

Number of clips	$L = 5$	$L = 15$	$L = 21$
mAP	89.18	89.69	89.74



(a) Single speaker

(b) Two speakers

(c) Background music

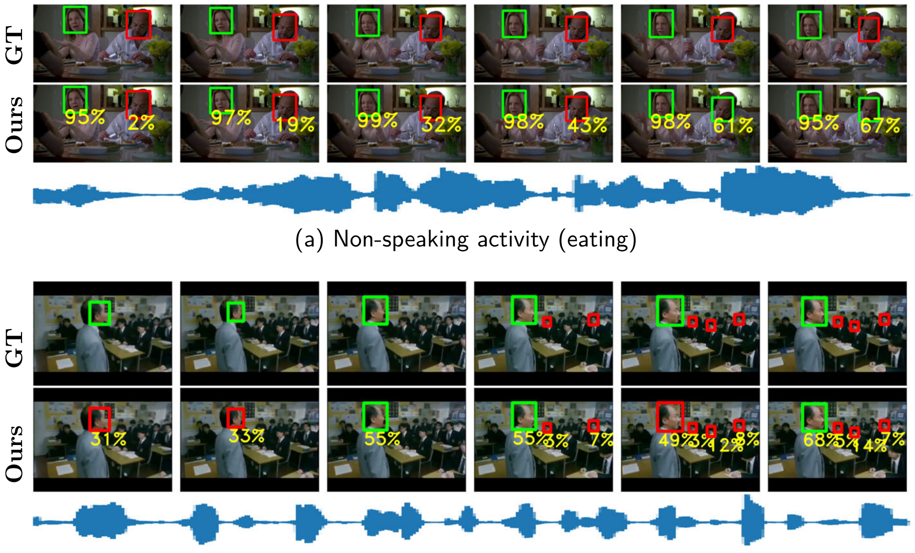
Fig. 4 Examples of successful results with AS-Net on the AVA-ActiveSpeaker dataset. (a) Single Speaker. (b) Two speakers. (c) Background music. The green and red boxes denote ground truth active speaker and non-active speaker, respectively. The speaking probability for each speaker is also provided. The time interval between adjacent frames is 500 ms

From the last line of Table 2, it is evident that enhancing the performance of AS-Net is possible, yielding an improvement of 0.44 mAP by fusing the outputs of the audio self-attention and audio-visual cross-attention networks. During the prototyping phase of the AS-Net model, we also experimented with audio-visual cross-attention between $\boldsymbol{F_a}$ and $\boldsymbol{F_v}$, concatenating the result with the output of the self-attention network, $\boldsymbol{F'_a}$. However, this alternative configuration did not yield a significant change in performance, as evidenced by a maintained mAP of 91.31, denoted as * in Table 2 and represented as * in Fig. 1.

4.4 Qualitative analysis

Figures 4 and 5 show a performance comparison between the AS-Net and ASC methods on short videos (3 s) featuring various scenarios from the AVA-ActiveSpeaker dataset. In Fig. 4a, it is evident that ASC struggles to differentiate speaking activities from non-speaking activities in the straightforward single-speaker scenario (columns 2, 4, and 5). In contrast, AS-Net's predictions of speaking probabilities align closely with the ground truths.

In scenarios featuring multiple speakers within a frame (Fig. 4b), AS-Net consistently maintains high speaking probabilities for active speakers and low probabilities for non-active speakers. On the other hand, ASC encounters challenges, particularly in predicting active speakers towards the end of speaking activities (see column 4). Moving to Fig. 4c, which depicts two speakers in the presence of background music where speech is inaudible, and the music noise is asynchronous with the speakers' facial movements. With this complex scenario, ASC produces false positive predictions by identifying music noise as speech (see column 2). Interestingly, AS-Net shows low speaking probabilities for both speakers, remaining unaffected by the background noise. This underscores the effectiveness of AS-Net in modeling audio-visual synchronization through the proposed attention schemes. Figure 4b



(a) Non-speaking activity (eating)

(b) Side view face

Fig. 5 Examples of ASD errors with AS-Net on the AVA-ActiveSpeaker dataset. (a) Non-speaking activity. (b) Side view face

and c prove that the proposed cross-attention scheme of AS-Net effectively connects the face movements with the dominant audio snippets.

On the contrary, Fig. 5 illustrates instances of ASD errors made by AS-Net. In Fig. 5a, a few false positive predictions are evident, specifically observed in the last two columns. This misdetection arises due to the presence of a non-active speaker eating while the active speaker engages in conversation with a third person in an open area, where the audio track may include both speech and background noise. Figure 5b presents a challenge as well, where side-view faces conceal facial movements during speech, leading AS-Net to struggle in predicting active speakers with high speaking probabilities (as seen in columns 1, 2, and 5). Despite this, the model avoids false negative predictions even in scenarios with multiple potential speakers in the same frame. However, the relatively low resolution of the speaker's face poses difficulties in extracting discriminative features for ASD.

A notable limitation stems from potential noise and unreliability in both audio and visual modalities, particularly in the audio domain. Real-world scenarios, such as those depicted in the AVA-ActiveSpeaker dataset, introduce non-speech sounds and strong background noise, posing a substantial challenge for achieving high accuracy in active speaker detection. To mitigate this limitation, there is a pressing need to suppress audio noises in the active speaker's surrounding environment or train AS-Net with non-speech audio samples mixed with the original clean speech to enhance its robustness in handling noisy data.

5 Conclusion

In this paper, a simple yet effective method for active speaker detection, AS-Net, was proposed. The idea behind AS-Net was to learn reliable spatial-temporal audio and visual embeddings by incorporating Temporal Shift Module (TSM) into 2D ResNet-18 and modeling long-term temporal audio-visual synchronization through simple self and cross-attention schemes. The efficiency of AS-Net in extracting robust spatial-temporal visual embeddings was demonstrated by integrating TSM and 2D ResNet-18, resulting in a notable 6.71 mAP improvement compared to using only the 2D ResNet-18. Moreover, AS-Net adeptly modeled long-term temporal synchronization between audio and visual modalities through its self-attention and cross-attention mechanisms, specifically designed to be cognizant of dominant audio and visual features. Experimental results demonstrated the superiority of AS-Net on the AVA-ActiveSpeaker dataset, outperforming a range of state-of-the-art 2D CNN-based methods with a 1.8 mAP advantage. In the future, our research will investigate the integration of information from inter-speakers, referring to speakers present in the same video frame as the target speaker, to serve as supplementary data for enhancing the target speaker's embedding.

Acknowledgements This work is supported by the Academy of Finland in project 345791. We acknowledge the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC and the LUMI consortium.

Funding Open Access funding provided by Aalto University.

Data Availability We have provided the references for the publicly available dataset used in the paper.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Chung S-W, Kang HG, Chung JS (2020) Seeing voices and hearing voices: learning discriminative embeddings using cross-modal self-supervision. In: INTERSPEECH. pp 3486–3490
2. Qian X, Brutti A, Lanz O, Omologo M, Cavallaro A (2021) Audio-visual tracking of concurrent speakers. IEEE Trans Multimed 24:942–954
3. Pibre L, Madrigal F, Equoy C, Lerasle F, Pellegrini T, Pinquier J, Ferrané I (2023) Audio-video fusion strategies for active speaker detection in meetings. Multimed Tools Appl 82(9):13667–13688
4. Wang Q, Downey C, Wan L, Mansfield PA, Moreno IL (2018) Speaker diarization with LSTM. In: ICASSP. pp 5239–5243
5. Cabañas-Molero P, Lucena M, Fuertes JM, Vera-Candeas P, Ruiz-Reyes N (2018) Multimodal speaker diarization for meetings using volume-evaluated SRP-PHAT and video analysis. Multimed Tools Appl 77:27685–27707
6. Chan DY, Wang J-F, Chin H-T (2023) A new speaker-diarization technology with denoising spectral-LSTM for online automatic multi-dialogue recording. Multimed Tools Appl 1–16
7. Chung JS, Zisserman A (2018) Learning to lip read words by watching videos. Comput Vision Image Understand 173:76–85
8. Kumar P, Malik S, Raman B (2023) Interpretable multimodal emotion recognition using hybrid fusion of speech and image data. Multimed Tools Appl 1–22
9. Chakravarty P, Mirzaei S, Tuytelaars T, Van hamme H (2015) Who's speaking? audio-supervised classification of active speakers in video. In: Proceedings of the 2015 ACM on international conference on multimodal interaction. pp 87–90
10. Fujita Y, Kanda N, Horiguchi S, Nagamatsu K, Watanabe S (2019) End-to-End Neural Speaker Diarization with Permutation-free Objectives. In: INTERSPEECH. pp 4300–4304
11. Stefanov K, Beskow J, Salvi G (2019) Self-supervised vision-based detection of the active speaker as support for socially aware language acquisition. IEEE Trans Cognit Develop Syst 12(2):250–259
12. Prajwal K, Afouras T, Zisserman A (2022) Sub-word level lip reading with visual attention. In: CVPR. pp 5162–5172
13. Chung JS, Zisserman A (2016) Out of time: automated lip sync in the wild. In: ACCV. pp 251–263
14. Gebru ID, Ba S, Li X, Horaud R (2017) Audio-visual speaker diarization based on spatiotemporal bayesian fusion. IEEE Trans Pattern Anal Mach Intell 40(5):1086–1099
15. Alcázar JL, Caba F, Mai L, Perazzi F, Lee J-Y, Arbeláez P, Ghanem B (2020) Active speakers in context. In: CVPR. pp 12465–12474
16. Roth J, Chaudhuri S, Klejch O, Marvin R, Gallagher A, Kaver L, Ramaswamy S, Stopczynski A, Schmid C, Xi Z et al (2020) Ava Active Speaker: An audio-visual dataset for active speaker detection. In: ICASSP. pp 4492–4496
17. Bulkin DA, Groh JM (2006) Seeing sounds: visual and auditory interactions in the brain. Curr Opinion Neurobiol 16(4):415–419
18. Ghaleb E, Niehues J, Asteriadis S (2023) Joint modelling of audio-visual cues using attention mechanisms for emotion recognition. Multimed Tools Appl 82(8):11239–11264
19. Alcázar JL, Caba F, Thabet AK, Ghanem B (2021) Maas: Multi-modal assignation for active speaker detection. In: ICCV. pp 265–274
20. Chung JS (2019) Naver at ActivityNet Challenge 2019–Task B Active Speaker Detection (AVA). arXiv:1906.10555
21. Carneiro H, Weber C, Wermter S (2021) FaVoA: Face-Voice association favours ambiguous speaker detection. In: ICANN. pp 439–450
22. Zhang Y, Liang S, Yang S, Liu X, Wu Z, Shan S, Chen X (2021) Unicon: Unified context network for robust active speaker detection. In: Proceedings of the 29th ACM international conference on multimedia. pp 3964–3972

23. Alcázar JL, Cordes M, Zhao C, Ghanem B (2022) End-to-end active speaker detection. In: ECCV. pp 126–143
24. Min K, Roy S, Tripathi S, Guha T, Majumdar S (2022) Learning long-term spatial-temporal graphs for active speaker detection. In: ECCV. pp 371–387
25. Köpüklü O, Taseska M, Rigoll G (2021) How to design a three-stage architecture for audio-visual active speaker detection in the wild. In: ICCV. pp 1193–1203
26. Huang C, Koishida K (2020) Improved active speaker detection based on optical flow. In: CVPR workshops. pp 950–951
27. Tao R, Pan Z, Das RK, Qian X, Shou MZ, Li H (2021) Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In: Proceedings of the 29th ACM international conference on multimedia. pp 3927–3935
28. Datta G, Etchart T, Yadav V, Hedau V, Natarajan P, Chang S-F (2022) Asd-Transformer: Efficient active speaker detection using self and multimodal transformers. In: ICASSP. pp 4568–4572
29. Xiong J, Zhou Y, Zhang P, Xie L, Huang W, Zha Y (2022) Look&listen: Multi-modal correlation learning for active speaker detection and speech enhancement. IEEE Trans Multimed 25:5800–5812
30. Medsker LR, Jain L (2001) Recurrent neural networks. Design Appl 5(64–67):2
31. Lin J, Gan C, Han S (2019) Tsm: Temporal shift module for efficient video understanding. In: ICCV. pp 7083–7093
32. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
33. Cho K, Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: EMNLP. pp 1724–1734
34. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
35. Tang H, Yuan C, Li Z, Tang J (2022) Learning attention-guided pyramidal features for few-shot fine-grained recognition. Pattern Recognit 130:108792
36. Tang H, Liu J, Yan S, Yan R, Li Z, Tang J (2023) M3net: Multi-view encoding, matching, and fusion for few-shot fine-grained action recognition. In: Proceedings of the 31st ACM international conference on multimedia. pp 1719–1728
37. Naik D, CD J (2023) Video captioning using sentence vector-enabled convolutional framework with short-connected LSTM. Multimed Tools Appl 1–27
38. Martin AF, Greenberg CS (2010) The NIST 2010 speaker recognition evaluation. In: INTERSPEECH. pp 2726–2729
39. Kim YJ, Heo H-S, Choe S, Chung S-W, Kwon Y, Lee B-J, Kwon Y, Chung JS (2021) Look who's talking: Active speaker detection in the wild. In: INTERSPEECH. pp 3675–3679
40. Zhang Y-H, Xiao J, Yang S, Shan S (2019) Multi-task learning for audio-visual active speaker detection. The ActivityNet Large-Scale Activity Recognition Challenge 1–4
41. Chung S-W, Chung JS, Kang H-G (2019) Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In: ICASSP. pp 3965–3969
42. Chatfield K, Simonyan K, Vedaldi A, Zisserman A (2014) Return of the devil in the details: Delving deep into convolutional nets. In: BMVC
43. Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45(11):2673–2681
44. Wang X, Cheng F, Bertasius G, Crandall D (2023) Loconet: Long-short context network for active speaker detection. arXiv:2301.08237
45. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR. pp 770–778
46. Kim C, Shin HV, Oh T-H, Kaspar A, Elgharib M, Matusik W (2018) On learning associations of faces and voices. In: ACCV. pp 276–292
47. Kingma DP, Ba J (2015) Adam: A method for stochastic optimizations. In: ICLR. pp 1–15