

ASDnB: Merging Face with Body Cues For Robust Active Speaker Detection

Anonymous IJCB 2025 submission

Abstract

State-of-the-art Active Speaker Detection (ASD) approaches mainly use audio and facial features as input. However, the main hypothesis in this paper is that body dynamics is also highly correlated to “speaking” (and “listening”) actions and should be particularly useful in wild conditions (e.g., surveillance settings), where face cannot be reliably accessed. We propose ASDnB, a model that singularly integrates face with body information by merging the inputs at different steps of feature extraction. Our approach splits 3D convolution into 2D and 1D to reduce computation cost without loss of performance, and is trained with adaptive weight feature importance for improved complement of face with body data. Our experiments show that ASDnB achieves state-of-the-art results in the benchmark dataset (AVA-ActiveSpeaker), in the challenging data of WASD, and in cross-domain settings using Columbia. This way, ASDnB can perform in multiple settings, which is positively regarded as a strong baseline for robust ASD models (code available at <https://github.com/anonymized>).

1. Introduction

Active Speaker Detection (ASD) aims to identify, from a set of potential candidates, active speakers on a given visual scene [33], with application in several topics such as speaker diarization [15, 10, 9], human-robot interaction [20, 40], automatic video editing [26, 14], and speaker tracking [30, 31]. State-of-the-art ASD models typically perform at the video frame level using face data and sound information. Using only facial cues as visual input is a viable strategy due to the correlation of mouth movement and speaking activity, but this approach is only reliable in cooperative and controlled settings. This widely used strategy is motivated by the benchmark ASD dataset AVA-ActiveSpeaker, composed of movies with good audio and face quality.

Another approach that has not been widely explored for ASD is the use of body information. When people speak or listen they typically use other forms of non-verbal be-

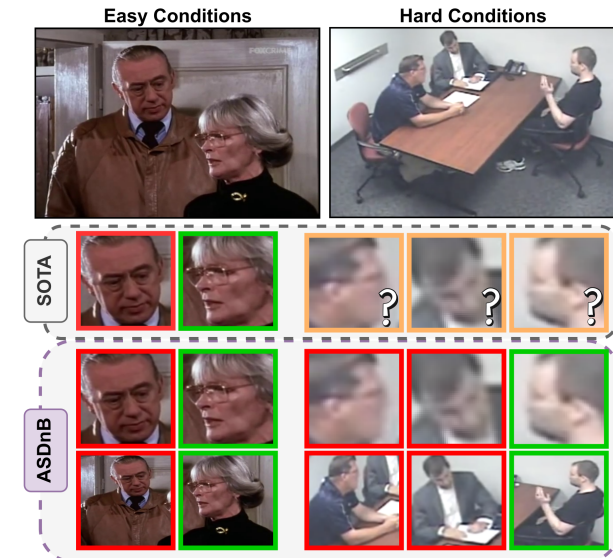


Figure 1. State-of-the-art ASD models solely rely on facial cues as visual input to perform. This approach is only reliable with cooperative (subjects) conditions, increasing uncertainty in model prediction in more challenging settings. Our approach aims to complement face with body cues to create more robust models that are able to perform in cooperative and unconstrained scenarios.

haviors such as head nodding, hand and body movements, which are not considered by current ASD models. This information increases in importance when face can not be reliably accessed (e.g., face occlusion) or in wilder conditions such as surveillance-like settings [35], as shown in Figure 1. As such, complementing body information with facial cues could improve model robustness to perform in cooperative conditions as well as more challenging settings.

This paper proposes a model that brings Active Speaker Detection and Body (ASDnB) together. In particular, ASDnB has the novelty of efficiently combining body with face data using a single visual encoder, merging them at different steps of the extraction, and outputting combined visual features for robust ASD. We modified the visual encoder by splitting the 3D convolution into 2D and 1D to reduce computation cost without loss of performance, and we train ASDnB using adaptive weight feature importance,

which results in improved visual encoder extraction and feature complement. Finally, we include temporal modeling in the classifier of ASDnB using bidirectional Gated Recurrent Unit (GRU) layers to maintain the temporal notion for speaker label prediction. Our experiments show that ASDnB achieves state-of-the-art results in the benchmark ASD dataset AVA-ActiveSpeaker, as well as in challenging data with degraded audio and face data quality (WASD [36]) and cross-domain settings (Columbia [6]), making it a baseline for robust ASD models. To summarize, the main contributions are:

- We announce the first effective combination of body and face data for visual input in ASD, which is a novel approach to create robust models to perform in more challenging settings;
- We propose a intra visual encoder combination of dual inputs (face and body) and training with adaptive weight feature importance to effectively combine relevant features for robust ASD;
- Ablation studies, experimental evaluation, and performance analysis demonstrate ASDnB state-of-the-art results in the benchmark dataset, AVA-ActiveSpeaker, in the challenging data of WASD, and in cross-domain settings using Columbia.

2. Related Work

ASD Context. Active Speaker Detection is the task to determine the talking speaker from a set of admissible candidates. The benchmark dataset of this area is AVA-ActiveSpeaker [33], which is based on Hollywood videos totalling almost 38 hours, with demographic diversity and FPS variation, with applications in other areas [34]. Several other datasets [3, 13, 21] were announced since, guaranteeing face access and good audio quality, similar to AVA-ActiveSpeaker, which is not an accurate representation of wild conditions [33]. For ASD in more challenging data, WASD [36] has been recently proposed containing different categories, with varying audio and face quality, ranging from cooperative conditions to surveillance settings. Based on the available data for ASD, current state-of-the-art models heavily rely on face and audio data, combining them using 3D architectures [8], hybrid 2D-3D models [48], and attention mechanisms [43, 1, 7]. Earlier works are based on a two-step process, where the first focuses on audio with face combination and the second on multi-speaker analysis [2, 22, 47, 3], while recently end-to-end models have emerged [42, 4, 29, 25, 35]. Contrary to existing works, where face is the only visual input for ASD, ASDnB is the first to effectively combine face with body data for robust ASD.

Model Enhancement. Strategies to improve ASD models are typically based on improved feature extraction and combination. Works focus on temporal speech refinement [2], inter-speaker and audio-visual relations [22, 47], using Graph Convolutional Networks (GCN) [45] to improve speaker relation representation [3, 29, 4], long-term temporal context with audio-visual synchronization [42], using a reference speech to improve ASD [19], and changing encoder architectures [25]. Despite the different strategies, combining audio with visual features is usually done post encoding using cross-attention approaches [42, 19] and complemented by temporal modeling. ASDnB is the first ASD model to combine different visual features (face and body), intra encoding, making it more robust to perform in challenging data where face can not be easily accessed.

Using Body Information. Current ASD state-of-the-art models rely on facial cues for visual input given the subject cooperation (face access guaranteed) of the mainstream datasets. However, this is not a viable approach in wilder conditions (such as surveillance settings), where face is not reliably accessed. As such, one potential strategy to improve ASD model robustness is using body information, as explored in other areas. Pedestrian Attribute Recognition (PAR) datasets [12, 28, 24] are examples of these scenarios, containing person cropped images from surveillance settings, used to identify various attributes under challenging covariates. Works in this area focus on different strategies ranging from different architecture combination [37, 49, 41], attention-based approaches [39, 17], assessing model limitations [38], and attribute relation importance [23, 27, 18]. In ASDnB, we propose a modification of standard ASD visual encoders, where face and body features are combined at different steps of the extraction, intra encoder, outputting combined visual features.

3. ASDnB

We propose ASDnB, an model that, for the first time, effectively combines face with body data to perform ASD in cooperative and challenging conditions. We combine dual visual inputs using a single encoder by complementing face and body features at different steps of extraction, and train with adaptive weight to improve feature extraction and combination. The overall architecture of ASDnB is displayed in Figure 2, with details of each model component in the following subsections.

3.1. Visual Encoder

Selecting Visual Encoding Approach. Several strategies for visual encoding in ASD are based on 3D convolutional neural networks, given their effectiveness in extracting spatiotemporal information of face sequences [22, 4]. These approaches are typically computationally expensive with increased number of parameters, which made other

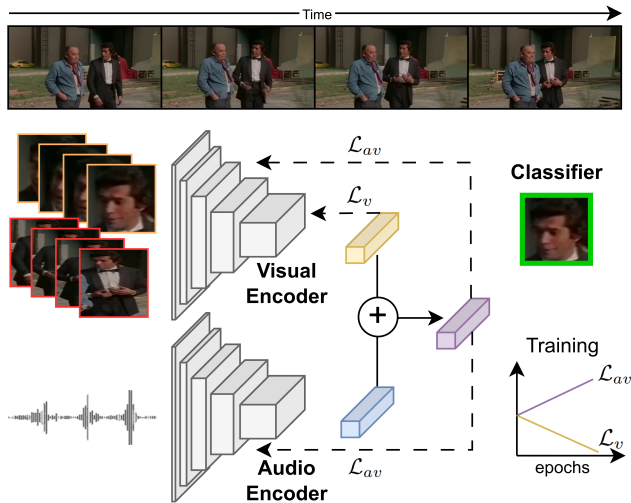


Figure 2. Overview of the ASDnB architecture. Body and face data are fed into the visual encoder, allowing intra encoder feature fusion and complement, while audio is processed through its respective encoder. Audio and visual features are combined to predict subject speaking label, using an adaptive weighted loss for combined and visual features (\mathcal{L}_{av} and \mathcal{L}_v , respectively).

works explore the use of a 3D convolution prior to inputting to a 2D ResNet (typically 18), followed by visual temporal convolution networks [42, 19]. The key takeaway from the state-of-the-art is that reducing the visual inputs to a 2D context leads to good ASD performance, which can be further extended by splitting the 3D convolution into 2D and 1D to extract the spatial and temporal information, respectively [25]. We select this approach for our model since it combines the key strategies of previous ASD works while significantly reducing the number of parameters and computational cost, and without loss of performance.

Combining Face and Body. Contrary to previous works, which only take face as input to visual encoders, we also consider body given its importance to complement facial cues for ASD in more challenging settings (*e.g.*, surveillance conditions). One possible approach is adding two visual encoders to extract face and body features, followed by combining them with audio prior to model classification. However, this does not force the model to consider face and body information in conjunction on feature extraction, resulting in ignoring information complement from the two sources, leading to subpar performances, namely in cooperative settings where face data is reliable [35]. As such, our motivation is to combine face and body data intra visual encoder to output combined visual feature, using an approach inspired by UNet [32] where face and body features are combined at different steps of the extraction, as shown in Figure 3. We combine face features into body at early extraction steps to make body data as a complement to facial cues since these are the main features for ASD,

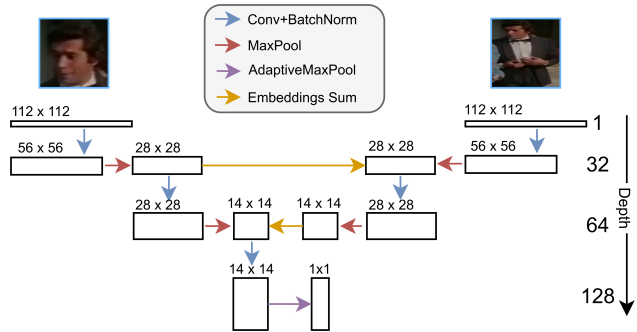


Figure 3. Overview of the flow of face and body combination in ASDnB visual encoder. The first convolution for both inputs downsamples via stride.

while recombining body data to face feature at latter stage to reinforce data conjunction. Finally, to ensure spatiotemporal feature extraction abundance at different receptive fields, each convolutional block contains two paths, one with kernel size of 3 and another of 5, given its superiority to other kernel combinations [25].

3.2. Audio Encoder

For audio encoding, we adapt the audio signal to serve as input to a 2D encoder by generating Mel-frequency cepstral coefficients (MFCCs), with a sampling rate of 16 kHz, analysis window of 0.025 ms, step between successive windows of 0.010, and a audio representation of 13 cepstrums. As audio encoder we use ResNet34 with Squeeze-and-Excitation (SE) blocks in its layers (SE-ResNet34), outputting a 128-dimensional audio embedding for subsequent visual and audio conjunction.

3.3. Temporal Modeling in Classifier

To improve ASD model performance we apply temporal modeling to the combined multi-modal features from audio and visual encoders. The key motivation is to provide a temporal notion to the model to predict if a subject is talking given that speaking is a continuous action in time, *i.e.* if one subject is talking in a given frame it is more likely to be talking in sequential frames, with a similar logic applying to a non-talking subject. Our approach for ASDnB is shown in Figure 4, where the combined multi-modal features are obtained by summing visual and audio features, followed by a bidirectional GRU, before passing to a Fully Connected (FC) layer to predict if the candidate is talking.

3.4. Loss Function

Selecting Loss Strategy. Existing ASD approaches are typically based on audio and face, with losses tending to rely on the conjunction of both data to assess if a subject is talking. However, recent works [42, 25] explore the inclusion of visual and audio features as sole inputs

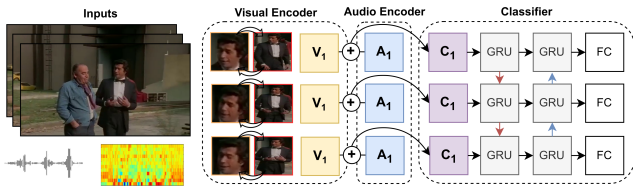


Figure 4. Bidirectional GRUs of visual and audio combination in ASDnB classifier, before inputting to FC layers for speaker classification.

for losses to complement the cross-modal interaction, leading to improved feature extraction and model performance. In our loss, we also consider visual features importance into it, given that cooperative ASD settings (like AVA-ActiveSpeaker) tend to benefit from assessing facial cues to predict subject talking (*i.e.*, mouth movement heavily relates to talking).

Feature Importance and Combination. Unlike previous works, we include the notion of gradually increment the importance of combined features (audio with visual) with a relative decrease of visual feature importance throughout training. The key motivation is that, although visual features are important to improve feature combination, its relevance is more crucial in earlier stages of training, with the end goal of ASD models being to assess if a subject is talking via both audio and visual, and not solely based on facial cues. This is particularly important for ASD in more challenging data (*e.g.*, surveillance settings), where the importance of visual features as sole input is not a reliable approach. Furthermore, since our visual encoder has to combine face and body features to output visual embeddings, we want the training to focus early on improving the visual encoder and later at combining audio with visual features.

Formally, we define ASD as a frame-level classification, where the predicted label sequence is compared with the ground truth via Cross-Entropy:

$$\mathcal{L} = -\frac{1}{T} \sum_{i=1}^T (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)), \quad (1)$$

where T refers to the number of video frames, p_i and y_i are the predicted and ground truth label for the i^{th} frame, respectively. Finally, the complete loss function is expressed by:

$$\mathcal{L}_{ASDnB} = \alpha \mathcal{L}_{av} + (1 - \alpha) \mathcal{L}_v, \quad (2)$$

where \mathcal{L}_{av} and \mathcal{L}_v refer to the losses of the combined features (audio with visual) and of the visual feature classification, respectively. α refers to the weight coefficients for combined features, with α starting at 0.5 and incrementing to 1 throughout training, as follows:

$$\alpha = \alpha_0 + \delta(\varepsilon - 1), \quad (3)$$

where α_0 is set to 0.5 as the initial coefficient importance, δ is set to $\frac{1}{60}$ as the coefficient decay degree, and ε refers to the training epoch.

3.5. Obtaining Body Data

One of our key contributions relates to combining body with facial cues to retrieve visual features relevant for ASD in varying conditions. However, most ASD datasets do not provide this type of data since current approaches rely solely on face information as visual input. Regarding AVA-ActiveSpeaker, we obtain body bounding box annotations from AVA Actions Dataset [16] (groundwork dataset) and complement them with speaking labels of AVA-ActiveSpeaker, by matching entity id of the original annotations. For Columbia, we use the S3FD face detector [46] based on previous works [25, 42], resizing its predictions to retrieve the body regions by using twice the width and ending the bottom of the bounding box at three times the predicted height. This approach is only viable for this setting given that subjects in Columbia are cropped to the upper body region (sitting). Regarding WASD [36], the original dataset already contains body data annotations.

3.6. Implementation Details

ASDnB is trained for 30 epochs with an Adam optimizer, with a initial learning rate of 10^{-4} , decreasing 5% for each epoch. All visual data is reshaped into 112×112 , audio data is represented by 13-dimensional MFCC, and both visual and audio features have an encoding dimension of 128. For visual augmentation, we perform random flip, rotate and crop, while for audio augmentation, we use negative audio sampling [42]. In sum, given a video data during training, a audio track of a new one is randomly selected from the same batch as noise, maintaining the same speaking label of the original soundtrack.

4. Experiments

4.1. Datasets and Evaluation Metrics

AVA-ActiveSpeaker. The AVA-ActiveSpeaker dataset [33] is an audio-visual active speaker dataset from Hollywood movies, ranging from 1 to 10 seconds, with 5.3 million face crops, where typically only train and validation sets are used for experiments [42, 25, 22, 19, 35].

WASD. The WASD dataset [36] compiles a set of videos from real interactions with varying accessibility of the two components for ASD: *audio* and *face*. With 30 hours of labelled data, WASD is divided into 5 categories ranging from optimal conditions to surveillance settings. We report the results on WASD and on each category, following WASD experiments.

Columbia. We also consider Columbia [6] following the methodology of Light-ASD [25] where models are trained

Table 1. Comparison of ASDnB with state-of-the-art models in AVA-ActiveSpeaker.

Model	Audio Encoder	Visual Encoder	Par(M)	Pre-training	End-to End	Body Data	mAP
ASC [2]	RN18 2D	RN18 2D	23.3	✓	×	×	87.1
MAAS [3]	RN18 2D	RN18 2D	21.7	✓	×	×	88.8
UniCon [47]	RN18 2D	RN18 2D	23.8	✓	×	×	92.2
TalkNet [42]	SE-RN34	RN18+V-TCN	15.0	×	✓	×	92.3
BIAS [35]	SE-RN34	RN18+V-TCN	31.6	×	✓	✓	92.4
ASD-Trans [11]	RN18 2D	RN18+V-TCN	15.0	×	✓	×	93.0
ASDNet [22]	SincDsNet	RNx101 3D	49.7	✓	×	×	93.5
TS-TalkNet [19]	SE-RN34	RN18+V-TCN	36.8	×	✓	×	93.9
EASEE-50 [5]	RN50	RN50 3D	74.7	✓	✓	×	94.1
Light-ASD [25]	Conv 1D	Conv 2D-1D	1.0	×	✓	×	94.1
SPELL [29]	RN18 2D	RN18+TSM	22.5	✓	×	×	94.2
ASDnB	SE-RN34	Conv 2D-1D	2.2	✓	✓	✓	94.6

in AVA-ActiveSpeaker, without any additional fine-tuning. Columbia consists of an 87-minute panel discussion video, with five speakers (Bell, Boll, Lieb, Long, and Sick) taking turns speaking, with 2-3 speakers visible at any given time.

Evaluation Metrics. For AVA-ActiveSpeaker and WASD, we use the official ActivityNet evaluation tool [33] that computes mean Average Precision (mAP), while for Columbia we use F1-Score.

4.2. ASDnB Performance in AVA-ActiveSpeaker

We compare ASDnB performance with state-of-the-art models in AVA-ActiveSpeaker, in Table 1. ASDnB outperforms other models while being **lightweight and trained end-to-end**. The increased number of parameters from other ASD approaches derive from heavier extraction power of visual inputs and model components (*e.g.*, GCN) to consider author relation which ASDnB simplifies by splitting 3D convolutions into 2D and 1D (lesser computation cost without loss of performance) and using temporal modeling in the classifier. However, the major contribution of ASDnB relative to existing approaches is the **first efficient combination of face with body data**. The inclusion of body information for ASD is extremely important, in particular for challenging data where face can not be reliably accessed, which is a novel strategy that state-of-the-art models do not yet consider. Only BIAS has previously considered complementing body with face data, but their approach was to treat body has another visual input rather than a complement to face leading to subpar results in AVA-ActiveSpeaker, where face is a reliable visual input for ASD. We are able to efficiently include body for ASD by combining face and body inputs intra encoder, outputting a single combined visual feature that complements facial cues with relevant body movements. Finally, our **pretraining strategy** is mainly to prepare ASDnB for adequate body information extraction by using WASD [36], a ASD dataset containing challenging data such as surveillance settings where face is not reliably

Table 2. Comparison of ASDnB with state-of-the-art models on the different categories of WASD, using the mAP metric. *OC* refers to Optimal Conditions, *SI* to Speech Impairment, *FO* to Face Occlusion, *HVN* to Human Voice Noise, and *SS* to Surveillance Settings. Light refers to Light-ASD, and TS-Talk to TS-TalkNet.

Model	Easy		Hard			Avg
	OC	SI	FO	HVN	SS	
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 [42]	95.8	97.5	93.1	81.4	77.5	92.3
TS-Talk [19]	96.8	97.9	94.4	84.0	79.3	93.1
Light [25]	97.8	98.3	95.4	84.7	77.9	93.7
BIAS [35]	97.8	98.4	95.9	85.6	82.5	94.5
ASDnB	98.7	98.9	97.2	89.5	82.7	95.6

accessed. Although AVA-ActiveSpeaker is the benchmark dataset, it is not a good representation of *in-the-wild* [33], with mainly cooperative conditions. As such, for body to have further importance in ASD we prepare the model to perform in harder settings prior to assess ASDnB in more cooperative conditions: AVA-ActiveSpeaker.

4.3. ASDnB Performance in Other Datasets

Challenging data of WASD. WASD is divided into categories with incremental challenges to audio and face data, with the most challenging data having face occlusion (FO), audio impairment with background voices (HVN) and surveillance settings (SS), where face access and audio quality is not guaranteed. Table 2 shows that ASDnB is superior to all models that only consider face as visual input, in particular for the Hard categories where face and audio quality is impacted, which rarely occurs in AVA-ActiveSpeaker. The biggest performance discrepancy is in surveillance settings (without reliable face input), where only BIAS performance is similar to ASDnB, given its strategy to also consider body data, highlighting the importance

Table 3. Comparison of F1-Score (%) on the Columbia dataset.

Model	Speaker					Avg
	Bell	Boll	Lieb	Long	Sick	
TalkNet [42]	43.6	66.6	68.7	43.8	58.1	56.2
LoCoNet [44]	54.0	49.1	80.2	80.4	76.8	68.1
Light-ASD [25]	82.7	75.7	87.0	74.5	85.4	81.1
BIAS [35]	89.3	75.4	92.1	88.8	88.6	86.8
ASDnB	91.6	81.2	93.1	91.7	90.6	89.6

Table 4. Ablation studies on the effect of WASD pretraining, face and body influence towards ASDnB performance (mAP) in AVA-ActiveSpeaker.

Face	Body	Pretrain	mAP
×	✓	×	83.9
		✓	86.5
✓	×	×	93.7
		✓	94.2
✓		×	94.1
	✓	✓	94.6

of including body information to create robust ASD models.

Robustness of ASDnB in Columbia. We also assess the performance of ASDnB in Columbia, following the methodology of Light-ASD [25] where models are trained in AVA-ActiveSpeaker, without any additional fine-tuning, and compare with the results reported on Light-ASD, in Table 3. Although Columbia data contains cooperative subjects, the cross-domain evaluation raises challenges for the models. In this context, ASDnB approach to combine body with face and audio information leads to a state-of-the-art performance, and highlights the relevance of complementing face with body data for model robustness to perform in varying conditions such as cross-domain settings.

4.4. Ablation Studies

Feature Influence. Given the ASDnB novelty of body inclusion for ASD, we explore the influence of different features, and pretraining in WASD, in Table 4. The main conclusions are: 1) the variant with only face as visual input and pretraining does not achieves state-of-the-art performances, meaning that body is a necessary feature; 2) pre-training benefits more the ASDnB variant with body than with only face (2.6% vs 0.5%), highlighting that pretraining in the challenges of WASD raises more importance to body information relative to facial cues; 3) the combination of body with face information, without pretraining, is the ASDnB variant with best performing results meaning that face with body complement is necessary for ASD but its relevance increases in more challenging data given that the best results require the WASD pretraining, which contains challenges not seen in AVA-ActiveSpeaker.

Feature Extraction. We explored variations of audio and visual encoders for ASDnB and summarized the re-

Table 5. Variation of audio and visual encoders regarding the number of parameters and model performance in AVA-ActiveSpeaker.

Visual Encoder	Audio Encoder	Par(M)	mAP
RN18+V-TCN	Conv 1D	16.5	92.5
RN18+V-TCN	SE-ResNet34	17.6	92.7
Conv 2D-1D	Conv 1D	1.1	94.1
Conv 2D-1D	SE-ResNet34	2.2	94.6

Table 6. Loss effect on ASDnB performance (mAP) in AVA-ActiveSpeaker. I_v refers to the importance of visual features, while I_{av} refers to the importance of combined audio and visual inputs towards ASD prediction. All approaches have Cross-Entropy as the underlying training loss. τ refers to the temperature coefficient.

Approach	I_v	I_{av}	Extra	mAP
Standard	0	1	×	93.1
TalkNet	0.4	1	×	94.0
Light-ASD	0.5	1	τ	94.2
Our	[0.5-0]	[0.5-1]	×	94.6

sults in Table 5. Regarding visual encoders, the approach of splitting 3D convolution into 2D and 1D to extract the spatial and temporal information, respectively, significantly outperforms the standard approach of using a ResNet with temporal convolutional network, while also having lower number of parameters. For ASD visual inputs, facial and body movements are the most relevant aspects, meaning that simpler models capture these notions better, without dispersion to other visual features. For audio encoder, the most robust approach of SE-ResNet34 leads to improved results when combined with the lightweight visual encoder. This is mainly due to the pretraining on WASD, with varying audio quality, that requires more robust extraction of audio features such as distinguish between relevant audio and background voices to combine with adequate visual information.

Loss Function. We compare our loss with approaches from other works in Table 6. Standard losses only consider the combined audio and visual features as relevant for ASD, while recent works [25, 42], focus on complementing this combined loss with weight importance of individual features. The latter strategies tend to perform better since they motivate ASD models to improve visual feature extraction such that visual cues are a reliable source to predict ASD. Light-ASD further improves this aspect by including a temperature coefficient to control feature importance throughout training epochs. Our approach is based on similar concepts but with two key changes: 1) our starting weight importance for visual and combined features increases the relative importance of visual features to motivate a better conjunction of face with body information in the earlier training stages; and 2) we vary visual and combined features impor-

Table 7. Performance of temporal modeling methods in ASDnB classifier.

Method	Par(M)	mAP
None	2.02	89.8
Forward LSTM	2.15	93.7
Forward GRU	2.12	93.8
Bidirectional LSTM	2.28	94.4
Bidirectional GRU	2.22	94.6

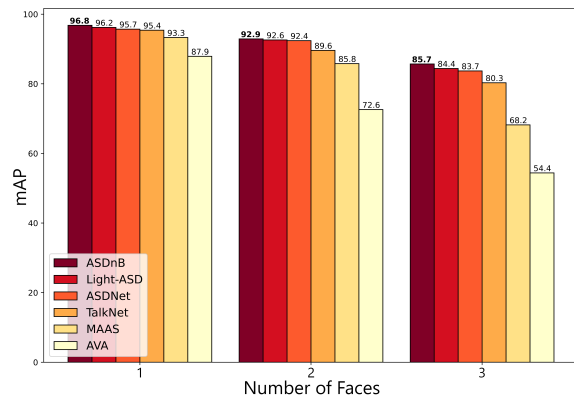
tance through training, such that in later stages visual features lose relevance and combining audio with visual features is the strategy for ASD. Our loss translates into better results relative to existing approaches, highlighting the influence of adaptive weight importance for more reliable ASD.

Temporal Modeling for ASD. We assess different temporal modeling approaches for ASDnB classifier, in Table 7. Given the ASD context, speaker prediction benefits from including a temporal relation between frames as shown by the results of not having temporal modeling. Increasing this relation by bidirectional (vs. forward) temporal modeling translates into better results, with GRU outperforming LSTM. LSTM tends to be more reliable for long-term information while the simplified version of GRU makes neighboring frames more informative, which is a better approach for ASD.

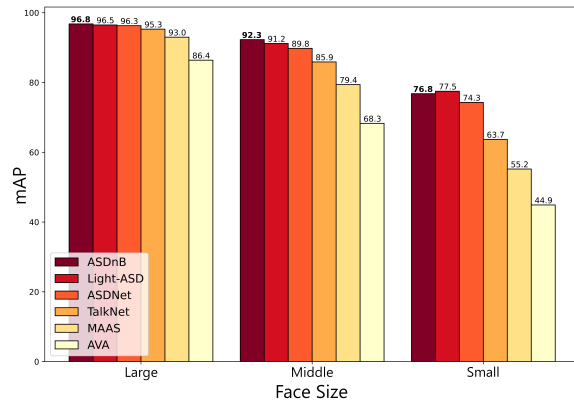
4.5. ASDnB Performance Analysis

Face Size and Number of People. We assess the robustness of ASDnB to deal with variations of AVA-ActiveSpeaker data in Figure 5, similar to other works [25, 42, 3, 22]. The methodology considers a face as *Small* with width under 64, *Middle* with width between 64 and 128, and *Large* with width over 128, while for the number of people in the scene the data is divided into three mutually exclusive groups (1, 2, and 3) based on the number of faces detected in a frame, totalling 90% of AVA-ActiveSpeaker data. For all variations ASDnB performance is superior to existing state-of-the-art models, with only a slight underperformance in the smaller face size relative to Light-ASD (76.8% vs. 77.5%). With smaller faces and no relevant body information ASDnB is not as robust, meaning that there is room for improvement in these settings, namely in background people of a scene. Aside this scenario, ASDnB is an all-around model for ASD, outperforming existing approaches in varying conditions.

Relative Body Importance. To further explore the importance of body for ASD, we compare the performance of ASDnB and Light-ASD with varying head-body proportion (HBP), in AVA-ActiveSpeaker, in Figure 7. We use all the available testing data in the first pair of bars, and use less data moving left to right on the x-axis, corresponding to lower HBP values. For instance, in the pair of bars at



(a) Models performance by the number of faces on each frame



(b) Models performance by faces size

Figure 5. Comparison of ASDnB performance relative to ASD state-of-the-art models for (a) number of faces per frame and (b) various face sizes in AVA-ActiveSpeaker.

20%, we are using the data with the lowest 20% HBP meaning that these are settings where the face is significantly smaller than body. The results show that, with a decrease of HBP, ASDnB performance is progressively better relative to Light-ASD, highlighting that the absence of reliable face access raises importance in body information. This is particularly predominant in wild conditions such as surveillance settings, meaning that ASDnB strategy is a viable approach to increase ASD robustness to perform in such conditions. Although ASDnB performs worse in smaller faces than Light-ASD (Figure 5), this aspect is mitigated with relevant body information meaning that the underperformance of ASDnB in such conditions is mainly due to background people without visible body. With relevant body information ASDnB is able to outperform Light-ASD, which reinforces the notion that body data is a relevant feature to use in challenging ASD data (WASD [36]), or when the subject is not close to the camera (instances of AVA-ActiveSpeaker).

Qualitative Analysis. We complement our experiments with a qualitative analysis of ASDnB and Light-ASD in challenging scenarios of AVA-ActiveSpeaker and WASD,

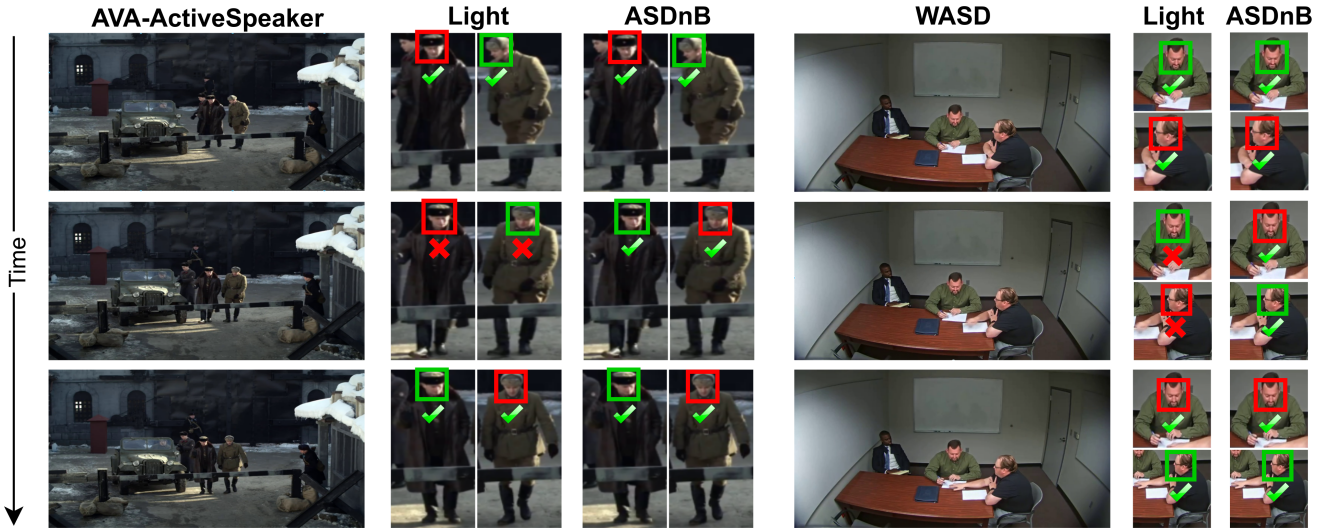


Figure 6. ASDnB and Light-ASD (Light) qualitative performance assessment in challenging scenarios of AVA-ActiveSpeaker and WASD. Red bounding boxes denote model prediction of subject not talking, green to speaking, and predictions with red cross denote misclassification relative to the ground-truth. In both examples with subjects far from the camera, Light-ASD misclassified the switch of speakers while ASDnB was more resilient by analysing the hand movement that preceded subject speaking.

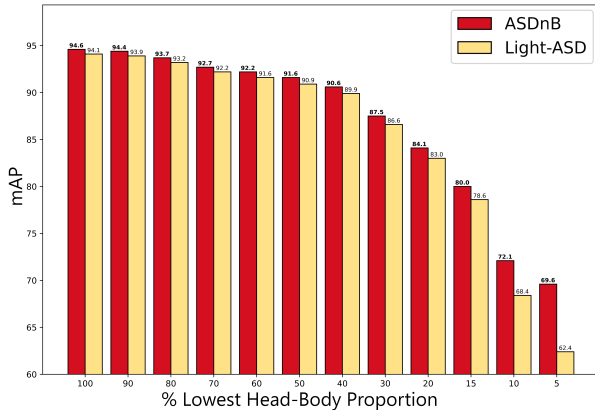


Figure 7. Relative body importance of ASDnB and Light-ASD with decremental head-body area proportion in AVA-ActiveSpeaker.

in Figure 6. The considered scenarios contain subjects far from the camera and in suboptimal cooperative settings (top-down view) which makes it harder to predict who is speaking using only facial cues. In both examples, Light-ASD misclassified the switch of speakers while ASDnB was more resilient by analysing the hand movement that preceded subject speaking. The results support the importance of body analysis for ASD in wild conditions, where face can not be reliably accessed, making ASDnB a viable baseline for robust ASD models.

5. Conclusion

This paper describes ASDnB, a lightweight multi-modal model that, for the first time, efficiently combines face

with body information for Active Speaker Detection. The key contribution of our proposal relates to combining face and body features at different feature extraction steps, inspired by the UNet approach, yielding state-of-the-art performance both on cooperative conditions (benchmark dataset AVA-ActiveSpeaker) and on more challenging settings (WASD and cross-domain of Columbia). The obtained results show that complementing body information with facial cues is of utmost importance for ASD robustness, and is particularly important for *wild* conditions (*i.e.*, surveillance settings), where state-of-the-art models do not reliably perform.

Acknowledgments

This work was supported by anonymous funding.

References

- [1] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman. Self-supervised learning of audio-visual objects from video. In *Proceedings of the ECCV*, pages 208–224. Springer, 2020.
- [2] J. L. Alcázar, F. Caba, L. Mai, F. Perazzi, J.-Y. Lee, P. Arbeláez, and B. Ghanem. Active speakers in context. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 12465–12474, 2020.
- [3] J. L. Alcázar, F. Caba, A. K. Thabet, and B. Ghanem. Maas: Multi-modal assignation for active speaker detection. In *Proceedings of the IEEE/CVF ICCV*, pages 265–274, 2021.
- [4] J. L. Alcazar, M. Cordes, C. Zhao, and B. Ghanem. End-to-end active speaker detection. *arXiv preprint arXiv:2203.14250*, 2022.
- [5] J. L. Alcazar, M. Cordes, C. Zhao, and B. Ghanem. End-to-end active speaker detection, 2022.

- [6] P. Chakravarty and T. Tuytelaars. Cross-modal supervision for learning active speaker detection in video. In *Proceedings of the ECCV*, pages 285–301. Springer, 2016.
- [7] Y. Cheng, R. Wang, Z. Pan, R. Feng, and Y. Zhang. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3884–3892, 2020.
- [8] J. S. Chung. Naver at activitynet challenge 2019–task b active speaker detection (ava). *arXiv preprint arXiv:1906.10555*, 2019.
- [9] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman. Spot the conversation: speaker diarisation in the wild. *Proc. Interspeech*, pages 299–303, 2020.
- [10] J. S. Chung, B.-J. Lee, and I. Han. Who said that?: Audio-visual speaker diarisation of real-world meetings. *Proc. Interspeech*, pages 371–375, 2019.
- [11] G. Datta, T. Etchart, V. Yadav, V. Hedau, P. Natarajan, and S.-F. Chang. Asd-transformer: Efficient active speaker detection using self and multimodal transformers. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4568–4572. IEEE, 2022.
- [12] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 789–792, 2014.
- [13] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra. Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments. *arXiv preprint arXiv:2107.04174*, 2021.
- [14] H. Duan, J. Liao, L. Lin, and W. Cai. Flad: a human-centered video content flaw detection system for meeting recordings. In *Proceedings of the 32nd Workshop on Network and Operating Systems Support for Digital Audio and Video*, pages 43–49, 2022.
- [15] I. D. Gebru, S. Ba, X. Li, and R. Horaud. Audio-visual speaker diarization based on spatiotemporal bayesian fusion. *IEEE TPAMI*, 40(5):1086–1099, 2017.
- [16] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018.
- [17] H. Guo, K. Zheng, X. Fan, H. Yu, and S. Wang. Visual attention consistency under image transforms for multi-label image classification. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 729–739, 2019.
- [18] J. Jia, H. Huang, W. Yang, X. Chen, and K. Huang. Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method. *arXiv preprint arXiv:2005.11909*, 2020.
- [19] Y. Jiang, R. Tao, Z. Pan, and H. Li. Target active speaker detection with audio-visual cues. In *Proc. Interspeech*, 2023.
- [20] S.-H. Kang and J.-H. Han. Video captioning based on both egocentric and exocentric views of robot vision for human-robot interaction. *International Journal of Social Robotics*, 15(4):631–641, 2023.
- [21] Y. J. Kim, H.-S. Heo, S. Choe, S.-W. Chung, Y. Kwon, B.-J. Lee, Y. Kwon, and J. S. Chung. Look who’s talking: Active speaker detection in the wild. *arXiv preprint arXiv:2108.07640*, 2021.
- [22] O. Köpüklü, M. Taseska, and G. Rigoll. How to design a three-stage architecture for audio-visual active speaker detection in the wild. In *Proceedings of the IEEE/CVF ICCV*, pages 1193–1203, 2021.
- [23] D. Li, X. Chen, and K. Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 111–115. IEEE, 2015.
- [24] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016.
- [25] J. Liao, H. Duan, K. Feng, W. Zhao, Y. Yang, and L. Chen. A light weight model for active speaker detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22932–22941, 2023.
- [26] J. Liao, H. Duan, X. Li, H. Xu, Y. Yang, W. Cai, Y. Chen, and L. Chen. Occlusion detection for automatic video editing. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2255–2263, 2020.
- [27] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95:151–161, 2019.
- [28] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE ICCV*, pages 350–359, 2017.
- [29] K. Min, S. Roy, S. Tripathi, T. Guha, and S. Majumdar. Learning long-term spatial-temporal graphs for active speaker detection. *arXiv preprint arXiv:2207.07783*, 2022.
- [30] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro. Audio-visual tracking of concurrent speakers. *IEEE Transactions on Multimedia*, 24:942–954, 2021.
- [31] X. Qian, M. Madhavi, Z. Pan, J. Wang, and H. Li. Multi-target doa estimation with an audio-visual fusion mechanism. In *2021 IEEE ICASSP*, pages 4280–4284. IEEE, 2021.
- [32] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.
- [33] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi, et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *2020 IEEE ICASSP*, pages 4492–4496. IEEE, 2020.
- [34] T. Roxo, J. C. Costa, P. R. Inácio, and H. Proença. On exploring audio anomaly in speech. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2023.

972		1026
973		1027
974	[35] T. Roxo, J. C. Costa, P. R. Inácio, and H. Proença. Bias: A body-based interpretable active speaker approach. <i>IEEE Transactions on Biometrics, Behavior, and Identity Science</i> , 2024.	1028
975		1029
976	[36] T. Roxo, J. C. Costa, P. R. M. Inácio, and H. Proença. Wasd: A wilder active speaker detection dataset. <i>IEEE Transactions on Biometrics, Behavior, and Identity Science</i> , pages 1–1, 2024.	1030
977		1031
978		1032
979		1033
980	[37] T. Roxo and H. Proença. Yinyang-net: Complementing face and body information for wild gender recognition. <i>IEEE Access</i> , 10:28122–28132, 2022.	1034
981		1035
982		1036
983	[38] T. Roxo and H. Proença. Is gender “in-the-wild” inference really a solved problem? <i>IEEE Transactions on Biometrics, Behavior, and Identity Science</i> , 3(4):573–582, 2021.	1037
984		1038
985		1039
986	[39] N. Sarafianos, X. Xu, and I. A. Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In <i>Proceedings of the ECCV</i> , pages 680–697, 2018.	1040
987		1041
988		1042
989	[40] G. Skantze. Turn-taking in conversational systems and human-robot interaction: a review. <i>Computer Speech & Language</i> , 67:101178, 2021.	1043
990		1044
991		1045
992	[41] C. Tang, L. Sheng, Z. Zhang, and X. Hu. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In <i>Proceedings of the IEEE ICCV</i> , pages 4997–5006, 2019.	1046
993		1047
994		1048
995	[42] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In <i>Proceedings of the 29th ACM International Conference on Multimedia</i> , pages 3927–3935, 2021.	1049
996		1050
997		1051
998		1052
999		1053
1000	[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. <i>Advances in neural information processing systems</i> , 30, 2017.	1054
1001		1055
1002		1056
1003		1057
1004	[44] X. Wang, F. Cheng, G. Bertasius, and D. Crandall. Loconet: Long-short context network for active speaker detection. <i>arXiv preprint arXiv:2301.08237</i> , 2023.	1058
1005		1059
1006		1060
1007	[45] M. Welling and T. N. Kipf. Semi-supervised classification with graph convolutional networks. In <i>Proceedings of the ICLR</i> , 2016.	1061
1008		1062
1009	[46] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S3fd: Single shot scale-invariant face detector. In <i>Proceedings of the IEEE international conference on computer vision</i> , pages 192–201, 2017.	1063
1010		1064
1011		1065
1012		1066
1013	[47] Y. Zhang, S. Liang, S. Yang, X. Liu, Z. Wu, S. Shan, and X. Chen. Unicon: Unified context network for robust active speaker detection. In <i>Proceedings of the 29th ACM International Conference on Multimedia</i> , pages 3964–3972, 2021.	1067
1014		1068
1015		1069
1016		1070
1017	[48] Y.-H. Zhang, J. Xiao, S. Yang, and S. Shan. Multi-task learning for audio-visual active speaker detection. <i>The ActivityNet Large-Scale Activity Recognition Challenge</i> , pages 1–4, 2019.	1071
1018		1072
1019		1073
1020	[49] X. Zhao, L. Sang, G. Ding, Y. Guo, and X. Jin. Grouping attribute recognition for pedestrian with joint recurrent learning. In <i>Proceedings of the IJCAI</i> , pages 3177–3183, 2018.	1074
1021		1075
1022		1076
1023		1077
1024		1078
1025		1079