WASD: A Wilder Active Speaker Detection Dataset

Tiago Roxo, Joana C. Costa, Pedro R. M. Inácio, *Senior Member, IEEE,* Hugo Proença, *Senior Member, IEEE*
Instituto de Telecomunicações, University of Beira Interior, Portugal
{tiago.roxo, joana.cabral.costa}@ubi.pt, {prmi, hugomcp}@di.ubi.pt

***Abstract*—Current Active Speaker Detection (ASD) models achieve good results on cooperative settings with reliable face access using only sound and facial features, which is not suited for less constrained conditions. To demonstrate this limitation of current datasets, we propose a Wilder Active Speaker Detection (WASD) dataset, with increased difficulty by targeting the key components of current ASD: audio and face. Grouped into 5 categories, WASD contains incremental challenges for ASD with tactical impairment of audio and face data, and provides a new source for ASD via subject body annotations. To highlight the new challenges of WASD, we divide it into Easy (cooperative settings) and Hard (audio and/or face are specifically degraded) groups, and assess state-of-the-art models performance in WASD and in the most challenging available ASD dataset: AVA-ActiveSpeaker. The results show that: 1) AVA-ActiveSpeaker prepares models for cooperative settings but not wilder ones (surveillance); and 2) current ASD approaches can not reliably perform in wilder settings, even if trained with challenging data. To prove the importance of body for wild ASD, we propose a baseline that complements body with face and audio information that surpass state-of-the-art models in WASD and Columbia. All contributions are available at https://github.com/Tiago-Roxo/WASD.**

***Index Terms*—Active speaker detection, body-based analysis, dataset, visual surveillance, wild conditions.**

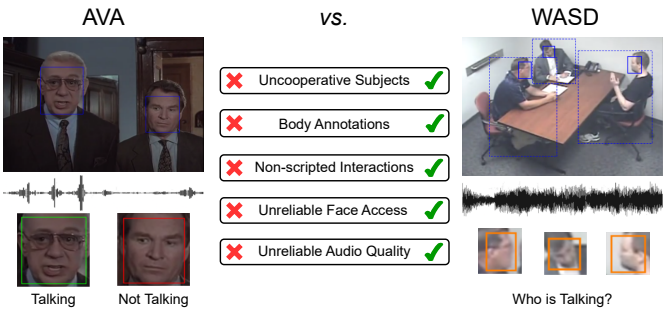


Fig. 1: AVA-ActiveSpeaker (AVA) state-of-the-art models achieve over 94% mean Average Precision (mAP) in active speaker detection, solely based on **face** and **audio data**. However, this approach may not be suited for uncooperative poses, non-guaranteed face access, or unreliable image/audio quality. How well do these models perform in such scenarios? And can body information aid in this task? To answer these questions (and more), we propose WASD, a Wilder Active Speaker Detection dataset.

I. Introduction

ACTIVE Speaker Detection (ASD) aims to identify, from a set of potential candidates, active speakers on a given visual scene [44]. Currently, this assessment is done at the video frame level using facial cues and sound information. Despite its application in several topics such as speaker diarization [12], [14], [25], human-robot interaction, or speaker tracking [40], [41], its applicability in wild conditions is still an open issue.

The state-of-the-art dataset for ASD is AVA-ActiveSpeaker [44], composed of several Hollywood movies, with diversity in languages, recording conditions, and speaker demographics, totalling in 38 hours and over 3 million face images. Although AVA-ActiveSpeaker has some challenging aspects, it still is not a perfect representation of *in-the-wild* data [44], since it assesses ASD in movies, a setup with controlled (scripted) action and speaking, with adequate audio and image quality. This motivates state-of-the-art models to identify active speakers solely based on audio and face data, disregarding other informations such as speaking context or body expressions. This is particularly problematic since ASD in wild conditions can not assume face availability, subject cooperation, and good audio quality, as shown in Figure 1. To overcome these limitations, we propose a Wilder Active Speaker Detection (WASD) Dataset.

Manuscript received XX, 2023; revised XX, 2023.

WASD aims to preserve the challenging characteristics of AVA-ActiveSpeaker while increasing the difficulty of ASD by targeting the two key components state-of-the-art models use: face and audio. We select videos from YouTube and group them into 5 categories, based on a set of features targeted at face and audio impairment. The categories range from optimal conditions (face availability and good audio quality), to surveillance settings (non-guaranteed face access, subject cooperation, or sound quality). The increasing scale of ASD challenges can be useful for: 1) assess the ability of current models to deal with wild conditions and specific aspect impairment (audio, face, or a combination of both); 2) evaluate the limitations of AVA-ActiveSpeaker to prepare models for wild conditions; and 3) show the limitations of face and audio dependency for wild ASD, easing the identification of model improvements towards this goal. By selecting YouTube videos from real interactions, WASD also contains expressions, sudden interruptions, and interactions that movies hardly contain. These additional challenges, enhanced by the variability of demographics in WASD, contribute to a challenging ASD dataset where state-of-the-art models can not easily perform. Furthermore, WASD provides body data annotations and a body-based approach to motivate the development of models using body information to complement face and audio data in (wild) ASD. Finally, the importance of WASD can be extended to other tasks, aside ASD, namely: speech recognition (what is the person saying), speech diarization (segment the audio per

TABLE I: Feature comparison of ASD datasets. AVA-ActiveSpeaker is represented as AVA. If datasets contain information regarding a feature, its absence is presented with ×, while its presence with ✓. WASD has a high number of hours, with increased number of faces and reduced face tracks (culminating in higher average video duration), Frames Per Second (FPS) variability, and increased talking percentage. The most disciminative factors are demographic representation, surveillance conditions, and body data annotations.

Dataset	Total Hours	Number of Faces (M)	Face Tracks (m)	Video Duration (s)	FPS Variability	Talking %	Demographic Representation	Surveillance Conditions	Body Data
Columbia [9]	1.5	0.2	-	-	×	-	×	×	✓
Talkies [5]	4.2	0.8	23.5	1.5	-	-	-	×	×
EasyCom [21]	6.0	-	-	-	×	-	×	×	×
ASW [32]	30.9	-	11.5	∼10	-	57.9	-	×	×
AVA [44]	37.9	3.7	38.5	≤10	✓	24.2	-	×	×
WASD	30.0	7.4	9.8	∼28	✓	84.6	✓	✓	✓

speaker), specific action recognition (dancing, talking, raising hand, social interactions), among others. To summarize, the main contributions are:

- We propose WASD, a ASD dataset divided into 5 categories with incremental ASD challenges, targeting audio quality and face availability, ranging from optimal conditions to surveillance settings. WASD is also innovative by providing ASD body annotations for various challenging conditions to motivate the development of body-based approaches;
- We show the limitations of AVA-ActiveSpeaker to prepare state-of-the-art ASD approaches for wilder conditions, and show that current ASD models can not reliably perform in such settings, even if trained in WASD data, in particular for audio impairment, facial occlusion, and surveillance settings;
- To prove the importance of body information for wild ASD, we propose a baseline that complements body with face and audio data, surpassing all state-of-the-art models, in particular for uncooperative and challenging sets (surveillance conditions) of WASD, and outperforms face-based implementations in Columbia.

II. Related Work

Active Speaker Detection. Works on ASD have evolved from facial visual cues [23], [38], [45] to audio as primary source [8], [20], to multi-modal data combination [4], [5], [33], [44], [49]. Since the introduction of AVA-ActiveSpeaker [44], combining audio with facial features is the *de facto* way to predict active speakers. Large 3D architectures [11], hybrid 2D-3D models [57], and large-scale pretraining [17], [19] for audio-visual combination are amongst some of the following works. Despite the viability of these approaches, feature embedding improvement [28] or attention approaches [3], [10], [50] were necessary to improve ASD. Creating two-step models, where the first focuses on short-term analysis (audio with face combination) and the second on multi-speaker analysis, is the approach from various recent works [4], [5], [33], [56]. ASC [4] focused on long-term multi-speaker analysis via temporal refinement, ASDNet [33] used a similar approach for inter-speaker relations, with improved visual backbones, and UniCon [56] relied on audio-visual relational contexts with various backbones. Improving speaker relation representation via Graph Convolutional Networks (GCN) [53] is also a viable approach to assess context information [5], [36]. Diverging from two-step training, end-to-end models have also emerged for ASD [6], [31], [35], [36], [49]. TalkNet [49] focused on improving long-term temporal context with audio-visual synchronization, TS-Talknet [31] considered pre-enrolled speaker embedding to complement this synchronization, EASEE [6] included GCN to complement spatial and temporal speaker relations, and Light-ASD [35] proposed a lightweight model by splitting 2D and 3D convolutions for audio-visual feature extraction, and applied Bidirectional Gated Recurrent Units (BGRU) for cross-modal modeling.

Datasets. There is a variety of available datasets suited for ASD, such as frontal speaker data, designed for speech recognition [29], [39], voice activity detection [48], and diarization [25] datasets. However, these are limited in subject diversity and talking scenarios, diminishing their relevance. With increased talking variability, datasets derived from movies and TV shows have also been reported [22], [26], [30], [43], limited by the low number of annotated hours. Other related setups are lip reading datasets [2], [15], [16], [18], [37], [47], whose purpose diverges from ASD since their goal is to infer the words pronounced from a given speaker. Recently there is a greater focus on specific ASD datasets [5], [9], [21], [32], [44], whose task is to determine the talking speaker from a set of admissible candidates. Columbia [9] contains 87 minutes of a panel discussion, with up to 3 visible speakers. Talkies [5] focuses on low duration videos, totalling 4 hours, with an average of 2.3 speakers and off-screen speaking. Easycom [21] is designed for multiple tasks related with augmented reality, composed of various sessions of speakers sat at a table, with background noise. AVA-ActiveSpeaker [44] is the state-of-the-art dataset, with over 150 Hollywood videos, totalling almost 38 hours, with demographic diversity and dubbed dialogues. ASW [32] was proposed with over 30 hours, from 212 videos randomly selected from the VoxConverse [13], containing various sets of interviews. The proposed dataset, WASD, brings challenging sets, *in-the-wild* videos, demographic diversity, and body data annotations. The main characteristics of our dataset relative to others are presented in Table I.

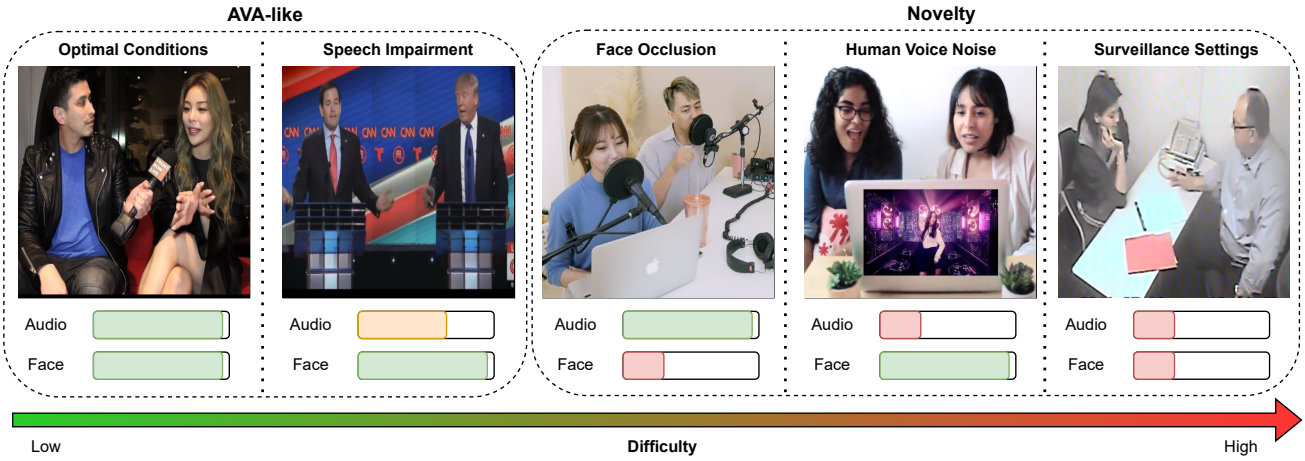


Fig. 2: Considered categories of WASD, with relative audio and face quality represented. Categories range from low (Optimal Conditions) to high (Surveillance Settings) ASD difficulty by varying audio and face quality. Easier categories contain similar characteristics to AVA-ActiveSpeaker (AVA-like), while harder ones are the novelty of WASD.

III. Dataset

We propose WASD, a dataset that aims to show the limitations of current state-of-the-art models by compiling a set of videos from real interactions with varying accessibility of the two key components for ASD: *audio* and *face*. By dividing our dataset into 5 categories with varying degrees of audio and face quality, we can assess how models adapt to these scenarios and which factors are more relevant for ASD. We create a balanced demographics dataset (regarding language, race, and gender), with several challenging factors, complemented with body annotations data. We discuss the process of dataset creation in the following sections.

A. Video and Category Selection

We select videos from YouTube and group them into 5 categories based on a set of features, whose values were attributed by human assessment. The main features used for category division are shown in Table II, with the complete list in the supplementary materials. In sum, videos are grouped as follows:

- **Optimal Conditions**: People talking in an alternate manner, with minor interruptions, cooperative poses, and face availability;
- **Speech Impairment**: Frontal pose subjects either talking via video conference call (*Delayed Speech*) or in a heated discussion, with potential talking overlap (*Speech Overlap*), but ensuring face availability;
- **Face Occlusion**: People talking with at least one of the subjects having partial facial occlusion, while keeping good speech quality (no delayed speech and minor communication overlap);
- **Human Voice Noise**: Communication between speakers where another human voice is playing in the background, with face availability and subject cooperation ensured;
- **Surveillance Settings**: Speaker communication in scenarios of video surveillance, with varying audio and image quality, without any guarantee of face access, speech quality, or subject cooperation.

TABLE II: Category feature matrix. Feature description: FA, Face Availability; SO, Speech Overlap; DS, Delayed Speech; FO, Facial Occlusion; HVB, Human Voice as Background Noise; SS, Surveillance Settings. The absence of a certain feature is presented with ×, while its presence with ✓. Features containing ? refer to non-guarantee of its presence or absence. Green cells refer to features favorable for ASD, while red ones are unfavorable.

Category	FA	SO	DS	FO	HVB	SS
Optimal Conditions	✓	×	×	×	×	×
Speech Impairment	✓	✓	✓	×	×	×
Face Occlusion	✓	×	×	✓	×	×
Human Voice Noise	✓	×	×	×	✓	×
Surveillance Settings	?	?	?	?	?	✓

Some important aspects to consider from Table II: 1) all categories, aside Surveillance Settings, guarantee face availability, which corresponds to cooperative scenarios and close-up faces; 2) we consider speech delay and overlap as variations of slight speech impairment, thus their grouping in the same category; and 3) Surveillance Settings does not have any guarantee regarding the analyzed features, corresponding to wild conditions. These considerations support the range of ASD difficulty between Optimal Conditions (easier) and Surveillance Settings (harder), since the impairment of audio and face is incremental and controlled throughout the categories. Figure 2 displays representative images of each category and the relative variation of audio and face quality. For a detailed list of all the considered features per video in WASD, we refer to the supplementary materials and the metadata Comma-Separated Values (CSV) containing this information, available on GitHub.

WASD Groups. Aside category division, we also form two groups of videos for our experiments: **Easy** and **Hard**. The easy group contains the categories that more closely

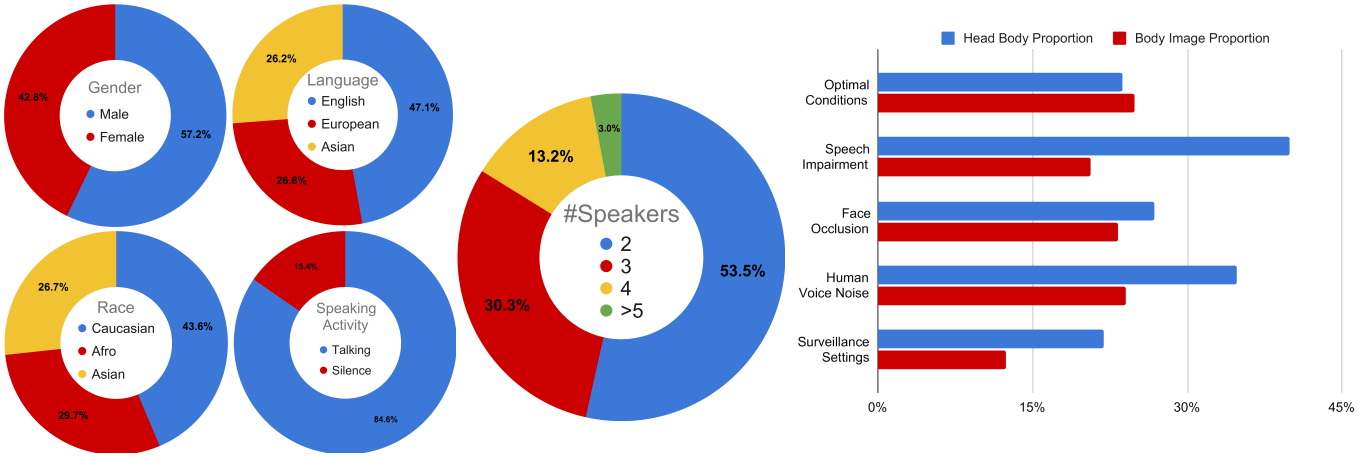


Fig. 3: Gender, language, race, speaking activity, and number of speakers distribution of WASD. Afro refers to African and Afro-American people. On the right, distribution of head-body and body-image proportions of WASD categories. WASD is a balanced demographics dataset, with *talking* being the predominant speaking activity, mainly composed of few people conversations, where audio impaired categories (Speech Impairment and Human Voice Noise) have speakers closer to the camera, and Surveillance Settings has speakers further from it.

resemble AVA-ActiveSpeaker (*Optimal Conditions* and *Speech Impairment*) while the hard group has categories where one or both factors (face and audio) are specifically degraded (remaining 3 categories of WASD). The inclusion of Speech Impairment in the easy group relates to how speech overlap is admissible in AVA-ActiveSpeaker (as recurrent from normal conversations) and speech delay as a result of dubbed movies (existent in AVA-ActiveSpeaker).

B. Main Characteristics

One focus of the proposed dataset is ensuring that each category is balanced regarding language, race, and gender distribution to mitigate any potential bias in future experiments. The languages are grouped into English, European, and Asian, while races are grouped into Caucasian, Afro, and Asian. The considered languages and races, their grouping, and other related considerations are discussed in the supplementary material. The distribution of demographics, number of speakers, and head-body proportions of WASD is presented in Figure 3. WASD only considers two admissible labels, with talking being the dominant speaking activity (contrary to AVA-ActiveSpeaker), and is mainly composed of few people conversations. Surveillance Settings is the one with lesser camera proximity to speakers while Speech Impairment and Human Voice Noise have speakers closer to the camera.

Following the AVA-ActiveSpeaker approach, the maximum length considered for each video is 15 minutes. Contrary to AVA-ActiveSpeaker, where each subvideo duration ranges up to 10 seconds, we segment each subvideo up to 30, with varying video FPS, mainly ranging from 24 to 30. Regarding the number of videos, WASD is composed of 164 videos (*vs.* 153 of AVA-ActiveSpeaker), totalling 30 hours of video annotations, divided into train and test with a similar proportion to AVA-ActiveSpeaker (80/20), with each category having the same amount of hours, (*i.e.*, 6 hours) and demographics balance.

C. WASD Annotations

Body bounding boxes drawing and tracking are obtained using YOLOv5 [42] and DeepSort [54], serving as input to Alphapose [24], [34], [55], which outputs pose information for each subject per frame. Then, we obtain face bounding boxes [7] from pose data, using eyes, ears, and nose keypoints as reference for bounding box drawing. The size of face bounding boxes is based on body bounding box height, which is adjusted manually per video to ensure adequate face capture. All face and body annotations are manually revised by a human and adjusted/fully annotated when necessary via Computer Vision Annotation Tool (CVAT) [46]. For speaking annotations, we design a custom Graphical User Interface (GUI) program in Python for manual annotation, outputting a file with the format used by AVA-ActiveSpeaker. Further details regarding annotations can be seen in the supplementary material.

IV. Experiments

A. Datasets, Models, and Evaluation Metric

Datasets. The AVA-ActiveSpeaker dataset [44] is an audio-visual active speaker dataset from Hollywood movies. With 262 15 minute videos, typically only train and validation sets are used for experiments: 120 for training, and 33 for validation, corresponding to 29,723 and 8,015 video utterances, respectively, ranging from 1 to 10 seconds. The main challenges of this dataset are related to language diversity, FPS variation, the existence of faces with low pixel numbers, blurry images, noisy audio, and dubbed dialogues. Similar to other works, we report the obtained results on the AVA-ActiveSpeaker validation subset. We also use the proposed dataset, WASD, which is described in Section III. Unless explicitly stated, all models trained in WASD use the whole training split (with 5 categories). We also consider Columbia [9] for cross-domain experiments with model modifications to include

TABLE III: Comparison of AVA-ActiveSpeaker trained state-of-the-art models on AVA-ActiveSpeaker and categories of WASD, using the mAP metric. We train and evaluate each model following the authors' implementation. *OC* refers to Optimal Conditions, *SI* to Speech Impairment, *FO* to Face Occlusion, *HVN* to Human Voice Noise, and *SS* to Surveillance Settings. AVA refers to AVA-ActiveSpeaker, Light refers to Light-ASD, and TS-Talk to TS-TalkNet.

Model	AVA	WASD					
		OC	SI	FO	HVN	SS	Avg
ASC [4]	83.6	86.4	84.8	69.9	66.4	51.1	74.6
MAAS [5]	82.0	83.3	81.3	68.6	65.6	46.0	70.7
ASDNet [33]	91.1	91.1	90.4	78.2	74.9	48.1	79.2
TalkNet [49]	91.8	91.6	93.0	86.4	77.2	64.6	85.0
TS-Talk [31]	92.7	91.1	93.7	88.6	79.2	64.0	85.7
Light [35]	93.4	93.1	93.8	88.7	80.1	65.2	86.2

TABLE IV: Comparison of 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 [4]	91.2	92.3	87.1	66.8	72.2	85.7
MAAS [5]	90.7	92.6	87.0	67.0	76.5	86.4
ASDNet [33]	96.5	97.4	92.1	77.4	77.8	92.0
TalkNet [49]	95.8	97.5	93.1	81.4	77.5	92.3
TS-Talk [31]	96.8	97.9	94.4	84.0	79.3	93.1
Light [35]	97.8	98.3	95.4	84.7	77.9	93.7

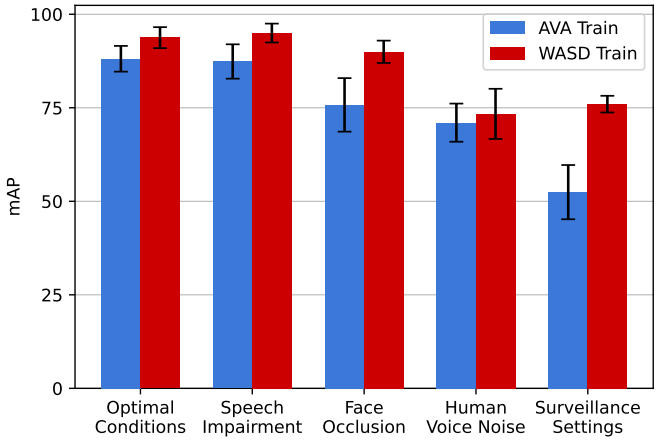


Fig. 4: Average performance (mAP) variation of the four models on WASD categories, when trained on AVA-ActiveSpeaker and WASD. AVA-ActiveSpeaker is represented as AVA.

body information. 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.

Models. The considered models are the ones with state-of-the-art results and publicly available implementations: ASC [4], MAAS [5], TalkNet [49], ASDNet [33], TS-TalkNet [31], and Light-ASD [35]. ASC, MAAS, and ASDNet are trained in a two-step process, while TalkNet, TS-TalkNet, and Light-ASD are trained end-to-end. MAAS did not provide its Multi-modal Graph Network setup so we present the results from the available implementation.

Evaluation Metric. For AVA-ActiveSpeaker and WASD, we use the official ActivityNet evaluation tool [44] that computes mean Average Precision (mAP), while for Columbia we use F1 score.

B. Limitations of AVA-ActiveSpeaker Training

We start by training models in AVA-ActiveSpeaker and evaluate their performance on AVA-ActiveSpeaker and WASD, in Table III. The inverse cross-domain (WASD to AVA-ActiveSpeaker) is included in the supplementary materials.

Similar to AVA-ActiveSpeaker. Regardless of the model, their performance on Easy categories (Optimal Conditions and Speech Impairment) is similar to the one displayed in AVA-ActiveSpeaker, suggesting the presence of similar characteristics between this group and AVA-ActiveSpeaker. This highlights the importance of face and audio quality for current ASD models, and shows that with high quality data and reliable face access, simultaneous talk or slight speech delay do not significantly hinder model performance. Furthermore, the similar performance of models in AVA-ActiveSpeaker and Easy categories support the quality of WASD annotations.

Face and Audio Importance. However, the cross-domain performance is significantly worse in Hard categories. In Face Occlusion, Human Voice Noise, and Surveillance Settings, there is a decrease in performance relative to other categories, suggesting that impairment of face access or audio quality significantly impact models, with a cumulative degrade when both are present (Surveillance Settings). Furthermore, facial occlusion is not as impactful as audio impairment (Human Voice Noise) in ASD, meaning that even when a model can not assess the talking person via face, it can still deduct it via audio analysis. The inverse is not as easily solved, since the existence of audio impairment with human voices (Human Voice Noise) leads to poorer performance relative to the Face Occlusion.

Model Performance Variance. One particular aspect regarding the performance of models for the Easy group is TalkNet and TS-TalkNet performing slightly better in Speech Impairment than in Optimal Conditions. This is mainly due to their long-term audio assessment approach, making the models robust against local audio variations and resilient to settings with slight audio impairment (Speech Impairment), even when trained in AVA-ActiveSpeaker, partially due to the existence of Speech Delay in AVA data (dubbed movies). Given this context, for TalkNet(s), Speech Impairment and Optimal Conditions tend to be "similar" environments, suggesting that the performance difference might be more related to the background of the videos and not their ASD challenges (*i.e.*, for TalkNet(s), Optimal Conditions and Speech Impairment only represent different videos and the latter is not necessarily harder). Additionally, Speech Impairment category has faces

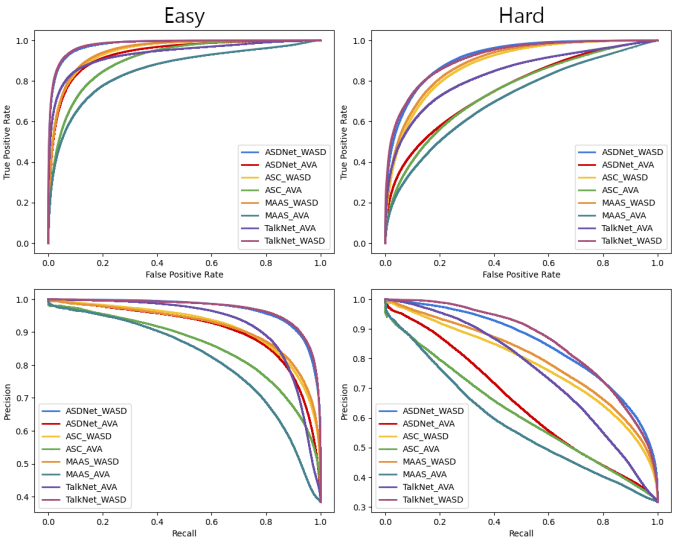


Fig. 5: ROC and PR curves for models trained in AVA-ActiveSpeaker and WASD, and evaluated in Easy (left) and Hard (right) groups of WASD. AVA-ActiveSpeaker is represented as AVA.

closer to the camera relative to Optimal Conditions (Figure 3, head-body proportion bar), which further contributes to the improvement of TalkNet(s) performance (*i.e.*, Speech Impairment is less challenging for TalkNet(s) than other models, and the face being closer to the camera leads to a slightly easier setting than Optimal Conditions).

Best Performers. Despite a performance degrade with increasing category difficulty, TalkNet, TS-TalkNet, and Light-ASD tend to perform better than the remaining models. This could be linked to their end-to-end approach for ASD, contrary to the other models, improving its generalization and performance in cross-domain. Furthermore, TalkNet and TS-TalkNet focus on long-term temporal context, benefiting from longer videos which is the case of WASD, while the BGRU for cross-modal modeling and lightweight model of Light-ASD contribute to more robustness for unseen conditions.

C. Models Robustness in WASD

To evaluate the robustness of ASD models on challenging data, we train them in WASD and compare their performance with AVA-ActiveSpeaker training in Table IV and Figure 4.

Performance Increase. Relative to AVA-ActiveSpeaker training, models trained in WASD tend to slightly improve their performance in Easy setups (Optimal Conditions and Speech Impairment), with higher increase in Face Occlusion and Surveillance Settings scenarios. The increase in Face Occlusion to closer values of those in Easy setups shows that, if trained accordingly, current models can perform ASD in such scenarios. This relates to how models can map different speaker relations in a scene, allowing the inference of one speaker relative to others, even if the face is occluded. Regarding Surveillance Settings, it shows that AVA-ActiveSpeaker does not contain data similar to these settings, but models can perform better in such scenarios if given the proper training.

Fig. 6: Incorrect model inference in different scenarios. Source of misconception: a) awe expression, with sudden and subtle mouth movement, while having human voice in the background; b) partial facial occlusion from scene object; and c) slight mouth occlusion from hand movement.

Similar to Face Occlusion, relating different speakers in a scene may give models the tools to perform in such scenarios, even when face access is not reliable.

Model Performance Differences. Regarding the performance of Human Voice Noise and Surveillance Settings, some models tend to perform better in the latter, even if said category is expected to be more challenging. We can look at the results considering two big groups: 1) ASDNet, TalkNet, Light-ASD, and TS-TalkNet (where Surveillance Settings is worse or equal to Human Voice Noise); and 2) MAAS and ASC (where Surveillance Settings is clearly better than Human Voice Noise). Older models (MAAS and ASC) are more heavily based on combining audio and face without attention-based approaches, and Human Voice Noise challenges an assumption of these models (given their training data and approaches): *audio of a given scene/video is very likely to come from ASD candidates* (*i.e.*, people in the scene). This is not guaranteed in this category since the majority of its videos have talking/singing audio from people not in the scene, which is distinctive from all other categories, such as Surveillance Settings (*i.e.*, in Surveillance Settings all talking audio comes from people in the scene), which might justify the underperformance of older models and the resilience of the most recent ones.

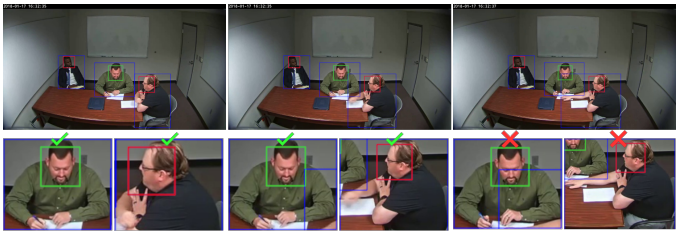


Fig. 7: Incorrect model inference by mixing active speakers. Hand and arm movement (from right speaker, better viewed with zoom in) suggest a change in conversation between the two speakers, whose analysis would aid understanding speaker swap mid conversation.

Model Limitations. When trained in WASD, models can not improve their performance in the presence of disruptive/distracting human voice background (Human Voice Noise), which shows the limitations of current approaches. The guaranteed face access may induce a false sense of security to classify a person as talking when they do micro expressions in the presence of (background) human voice. Furthermore, the disparity between the results with human voice background or surveillance settings and the other scenarios (77% *vs* >92%) shows the limitations of current models to perform in wilder ASD contexts, particularly in impaired audio conditions.

Performance in WASD Groups. To complement model performance assessment, we compute the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves of models in different experimental settings, in Figure 5. The results show that: 1) in the Easy group, ASDNet and TalkNet trained in AVA-ActiveSpeaker are competitive with other models trained in WASD, showing the robustness of the best performing models and the similarity between AVA-ActiveSpeaker and Easy group of WASD; 2) for the Hard group, all models trained in WASD have superior performance relative to AVA-ActiveSpeaker training, suggesting the difference of data between this group and AVA-ActiveSpeaker; and 3) TalkNet trained in AVA-ActiveSpeaker displays a different tendency relative to other models, expressed in both Easy and Hard group, with higher predominance in PR curves. TalkNet has a cautious and precise approach in determining the active speaker (high precision), while not keeping a similar performance in identifying all the active speakers as other models (lower precision with higher recall). This is linked to the lower talking percentage of AVA-ActiveSpeaker and the end-to-end approach of TalkNet with emphasis on long term context: identifying only active speakers with high confidence is a good strategy in AVA-ActiveSpeaker but not as reliable in WASD.

D. Qualitative Analysis

We analyze different scenarios where WASD is distinctive from AVA-ActiveSpeaker and body data analysis is more relevant for ASD, namely in Human Voice Noise, Face Occlusion, and Surveillance Settings, represented in Figures 6a, 6b, 6c, and 7, respectively. Head boxes are colored with models predictions, trained in WASD: green, person is talking; red, not talking. Figures are accompanied with zoom ins containing wrong and correct signs, displaying the correctness of ASD prediction. By not using body information, state-of-the-art models can not reliably deal with scenarios where someone expresses slight lip movement (*e.g.*, awe expression) when another person (not in scene) is talking (Figure 6a), or with facial occlusion (Figure 6b), even in the context of speaker proximity and cooperation. In surveillance settings (Figures 6c and 7) the benefits of body data evaluation is even more pronounced. Accessing hand movement with slight face occlusion helps understanding that the same person is talking (Figure 6c), as well as inferring when one person is requesting other to stop talking (Figure 7).

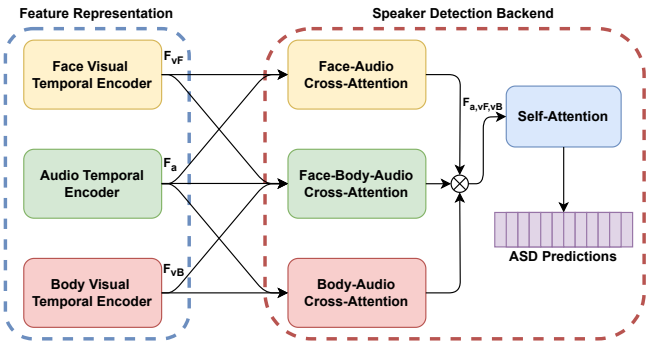


Fig. 8: Adaptation of TalkNet [49] architecture to include body data and serve as our baseline: inclusion of a visual temporal encoder for body data and cross-attention for audio and body.

V. Proposed Baseline

Given the superiority of TalkNet in WASD and its end-to-end approach, we select it as our baseline, modifying its architecture to include body information while maintaining its key characteristics, as shown in Figure 8. In experiments with only one visual data input (face or body), we use the original TalkNet.

A. TalkNet Architecture Modifications

To ensure adequate body data encoding, we add a new visual encoder for body data, with the same architecture as the one used for face (V-TCN [49] with DS-Conv1D [1]). For multi-modal cross-attention, we combine audio and body features following the approach used for face and audio [51]. Feature combination was adapted to include body and self-attention block input was updated to consider these changes. Given that the original training loss function considers the prediction of feature combination and each feature individually with weighted factors, we include a body loss with the same weight as face.

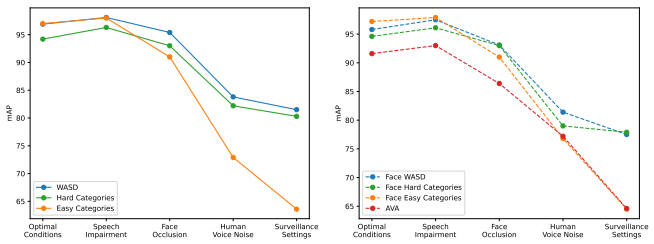
B. Body Importance Assessment

We assess the performance of the proposed Baseline and TalkNet using face and body data, in Table V.

Body Data Alone. Although body data contains facial cues, model attention to them diminishes when using body

TABLE V: Comparison of TalkNet, using face and body data, and proposed Baseline on WASD categories, 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.

Model	Easy		Hard		
	OC	SI	FO	HVN	SS
TalkNet$_{Face}$	95.8	97.5	93.1	81.4	77.5
TalkNet$_{Body}$	91.1	95.5	88.4	73.1	75.0
Baseline	96.9	98.1	95.4	83.8	81.5



(a) (b)

Fig. 9: WASD group training performance (mAP) influence on: a) Proposed Baseline; and b) TalkNet with face as input. AVA refers to TalkNet trained in AVA-ActiveSpeaker, and WASD refers to training using all training data.

as sole input, leading to worse performance overall. The categories less affected by this change are Speech Impairment and Surveillance Settings, given the similarity of face and body data (head body proportion in Table 3), and reduced face importance due to its unreliable access, respectively. This shows that current ASD approaches can not effectively process body data as input and are more prepared to analyze closed-up facial/mouth cues for ASD prediction.

Face Body Importance Discrepancy. Regarding the difference between using face or body alone, Human Voice Noise is the category where this difference is maxed. Human Voice Noise has videos with sporadic/involuntary body movements that might suggest that a person is talking, while not actually being (*e.g.*, dancing, pointing at something, raising hand in reaction to a given event). In these cases, considering only body information as visual input in Human Voice Noise data which contains movements that are rarely/never seen in other categories, might be detrimental towards ASD and lead to false positives based on sudden/abrupt body movements (*i.e.*, when looking at other categories, raising hands or similar actions are heavily related to speaking, which is not guaranteed in Human Voice Noise).

Face Body Complement. The combination of body with face data translates into better results for all categories, with the biggest improvement on Hard categories. This shows that face-body complement is a viable strategy for current models and body importance is emphasized in harder setups, particularly when face can not be reliably accessed.

TABLE VI: Comparison of F1-Score (%) on the Columbia dataset.

Model	Speaker					
	Bell	Boll	Lieb	Long	Sick	Avg
TalkNet [49]	43.6	66.6	68.7	43.8	58.1	56.2
LoCoNet [52]	54.0	49.1	80.2	80.4	76.8	68.1
Light-ASD [35]	82.7	**75.7**	**87.0**	74.5	**85.4**	**81.1**
Baseline	**83.0**	64.5	79.6	**86.2**	78.9	78.4

C. Training Influence in Model Performance

We explore the effect of training in WASD groups (Easy and Hard, Section III-A) on the proposed Baseline and TalkNet with face as visual input, in Figure 9.

Group Training on Baseline. For our Baseline, training in Easy has similar performance to training in WASD for cooperative setups, but significantly underperforms in unseen challenging sets. In contrast, Hard training follows the trend of WASD training (with only slightly worse performance), showing the importance of training in challenging data for robustness in within and cross-domain settings.

Training Influence on Face-based Models. Similar to our Baseline, face-based models overall performance is better when training in Hard than in Easy. However, while in Baseline there is a gap between Hard and WASD training, in face-based models the performance is more similar, showing the importance of having additional training data for more robust approaches (body-face models). Relative to TalkNet training in AVA-ActiveSpeaker, its performance on WASD follows the same trend as Easy training, reinforcing that AVA-ActiveSpeaker and Easy data are similar, and that neither adequately prepare models to deal with challenging sets.

D. Body Importance in Columbia

We also assess the performance of the proposed baseline in Columbia, following the procedure of Light-ASD [35] where models are trained in AVA-ActiveSpeaker, without any additional fine-tuning, and compare with the results reported on Light-ASD, in Table VI. Since AVA-ActiveSpeaker does not have body data annotations, we obtain body bounding box annotations from AVA Actions Dataset [27] and complement them with speaking labels of AVA-ActiveSpeaker to train the proposed baseline. The results show that body inclusion massively improves the proposed baseline performance relative to its initial architecture (TalkNet), showing that body information leads to increased robustness to perform ASD and can be useful in other settings other than the challenging conditions of WASD.

VI. Conclusion

We propose WASD, a innovative challenging ASD dataset with degraded audio quality, facial occlusions, surveillance conditions, and body data annotations. We show that current datasets do not prepare models for wild ASD, and state-of-the-art models underperform in such conditions, particularly in audio impairment and surveillance settings. We also propose a baseline that complements body data with audio and face to

prove the importance of body for wild ASD and support the development of subsequent approaches, given the unreliability of audio quality and subject cooperation in wilder settings. Given the challenges of the surveillance data relative to existing datasets, future directions should be centered on increasing the amount of challenging data available to prepare models to perform in wilder conditions.

ACKNOWLEDGMENTS

This work was supported by the Project UIDB/50008/2020, FCT Doctoral Grants 2020.09847.BD and 2021.04905.BD, and Project CENTRO-01-0145-FEDER-000019.

REFERENCES

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727, 2018. 7

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 2

[3] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision*, pages 208–224. Springer, 2020. 2

[4] Juan León Alcázar, Fabian Caba, Long Mai, Federico Perazzi, Joon-Young Lee, Pablo Arbeláez, and Bernard Ghanem. Active speakers in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12465–12474, 2020. 2, 5

[5] Juan Léon Alcázar, Fabian Caba, Ali K Thabet, and Bernard Ghanem. Maas: Multi-modal assignation for active speaker detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 265–274, 2021. 2, 5

[6] Juan Leon Alcazar, Moritz Cordes, Chen Zhao, and Bernard Ghanem. End-to-end active speaker detection. *arXiv preprint arXiv:2203.14250*, 2022. 2

[7] Anonymous Authors. Anonymous title. *Anonymous Journal/Conference*, page Anonymous Pages, Anonymous Year. 4

[8] Punarjay Chakravarty, Sayeh Mirzaei, Tinne Tuytelaars, and Hugo Van hamme. Who's speaking? audio-supervised classification of active speakers in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 87–90, 2015. 2

[9] Punarjay Chakravarty and Tinne Tuytelaars. Cross-modal supervision for learning active speaker detection in video. In *European Conference on Computer Vision*, pages 285–301. Springer, 2016. 2, 4

[10] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie 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. 2

[11] Joon Son Chung. Naver at activitynet challenge 2019–task b active speaker detection (ava). *arXiv preprint arXiv:1906.10555*, 2019. 2

[12] Joon Son Chung, Jaesung Huh, Arsha Nagrani, Triantafyllos Afouras, and Andrew Zisserman. Spot the conversation: speaker diarisation in the wild. *Proc. Interspeech*, pages 299–303, 2020. 1

[13] Joon Son Chung, Jaesung Huh, Arsha Nagrani, Triantafyllos Afouras, and Andrew Zisserman. Spot the Conversation: Speaker Diarisation in the Wild. In *Proc. Interspeech 2020*, pages 299–303, 2020. 2

[14] Joon Son Chung, Bong-Jin Lee, and Icksang Han. Who said that?: Audio-visual speaker diarisation of real-world meetings. *Proc. Interspeech*, pages 371–375, 2019. 1

[15] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *Proc. Interspeech 2018*, pages 1086–1090, 2018. 2

[16] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian conference on computer vision*, pages 87–103. Springer, 2016. 2

[17] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 2

[18] Joon Son Chung and AP Zisserman. Lip reading in profile. *British Machine Vision Conference (BMVC)*, 2017. 2

[19] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3965–3969. IEEE, 2019. 2

[20] Shaojin Ding, Quan Wang, Shuo-Yiin Chang, Li Wan, and Ignacio-Lopez Moreno. Personal vad: Speaker-conditioned voice activity detection. In *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, pages 433–439, 2020. 2

[21] Jacob Donley, Vladimir Tourbabin, Jung-Suk Lee, Mark Broyles, Hao Jiang, Jie Shen, Maja Pantic, Vamsi Krishna Ithapu, and Ravish Mehra. Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments. *arXiv preprint arXiv:2107.04174*, 2021. 2

[22] Mark Everingham, Josef Sivic, and Andrew Zisserman. Hello! my name is... buffy"–automatic naming of characters in tv video. In *BMVC*, volume 2, page 6, 2006. 2

[23] Mark Everingham, Josef Sivic, and Andrew Zisserman. Taking the bite out of automated naming of characters in tv video. *Image and Vision Computing*, 27(5):545–559, 2009. 2

[24] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 4

[25] Israel D Gebru, Sileye Ba, Xiaofei Li, and Radu Horaud. Audio-visual speaker diarization based on spatiotemporal bayesian fusion. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1086–1099, 2017. 1, 2

[26] Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard. The repere corpus: a multimodal corpus for person recognition. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1102–1107, 2012. 2

[27] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul 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. 8

[28] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 2

[29] Timothy J Hazen, Kate Saenko, Chia-Hao La, and James R Glass. A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 235–242, 2004. 2

[30] Yongtao Hu, Jimmy SJ Ren, Jingwen Dai, Chang Yuan, Li Xu, and Wenping Wang. Deep multimodal speaker naming. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1107–1110, 2015. 2

[31] Yidi Jiang, Ruijie Tao, Zexu Pan, and Haizhou Li. Target active speaker detection with audio-visual cues. In *Proc. Interspeech*, 2023. 2, 5

[32] You Jin Kim, Hee-Soo Heo, Soyeon Choe, Soo-Whan Chung, Yoohwan Kwon, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. Look who's talking: Active speaker detection in the wild. *arXiv preprint arXiv:2108.07640*, 2021. 2

[33] Okan Köpüklü, Maja Taseska, and Gerhard Rigoll. How to design a three-stage architecture for audio-visual active speaker detection in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1193–1203, 2021. 2, 5

[34] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*, 2018. 4

[35] Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, and Liangyin 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. 2, 5, 8

[36] Kyle Min, Sourya Roy, Subarna Tripathi, Tanaya Guha, and Somdeb Majumdar. Learning long-term spatial-temporal graphs for active speaker detection. *arXiv preprint arXiv:2207.07783*, 2022. 2

[37] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. In *Proc. Interspeech 2017*, pages 2616–2620, 2017. 2

[38] Foteini Patrona, Alexandros Iosifidis, Anastasios Tefas, Nikolaos Nikolaidis, and Ioannis Pitas. Visual voice activity detection in the wild. *IEEE Transactions on Multimedia*, 18(6):967–977, 2016. 2

[39] Eric K Patterson, Sabri Gurbuz, Zekeriya Tufekci, and John N Gowdy. Cuave: A new audio-visual database for multimodal human-computer interface research. In *2002 IEEE International conference on acoustics, speech, and signal processing*, volume 2, pages II–2017. IEEE, 2002. 2

[40] Xinyuan Qian, Alessio Brutti, Oswald Lanz, Maurizio Omologo, and Andrea Cavallaro. Audio-visual tracking of concurrent speakers. *IEEE Transactions on Multimedia*, 24:942–954, 2021. 1

[41] Xinyuan Qian, Maulik Madhavi, Zexu Pan, Jiadong Wang, and Haizhou Li. Multi-target doa estimation with an audio-visual fusion mechanism.

In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4280–4284. IEEE, 2021. 1

[42] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 4

[43] Jimmy Ren, Yongtao Hu, Yu-Wing Tai, Chuan Wang, Li Xu, Wenxiu Sun, and Qiong Yan. Look, listen and learn—a multimodal lstm for speaker identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016. 2

[44] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4492–4496. IEEE, 2020. 1, 2, 4, 5

[45] Kate Saenko, Karen Livescu, Michael Siracusa, Kevin Wilson, James Glass, and Trevor Darrell. Visual speech recognition with loosely synchronized feature streams. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1424–1431. IEEE, 2005. 2

[46] Boris Sekachev, Nikita Manovich, Maxim Zhiltsov, Andrey Zhavoronkov, Dmitry Kalinin, Ben Hoff, TOsmanov, Dmitry Kruchinin, Artyom Zankevich, DmitriySidnev, Maksim Markelov, Johannes222, Mathis Chenuet, a andre, telenachos, Aleksandr Melnikov, Jijoong Kim, Liron Ilouz, Nikita Glazov, Priya4607, Rush Tehrani, Seungwon Jeong, Vladimir Skubriev, Sebastian Yonekura, vugia truong, zliang7, lizhming, and Tritin Truong. opencv/cvat: v1.1.0, Aug. 2020. 4

[47] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6447–6456, 2017. 2

[48] Fei Tao and Carlos Busso. Bimodal recurrent neural network for audiovisual voice activity detection. In *INTERSPEECH*, pages 1938–1942, 2017. 2

[49] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3927–3935, 2021. 2, 5, 7, 8

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 7

[52] Xizi Wang, Feng Cheng, Gedas Bertasius, and David Crandall. Loconet: Long-short context network for active speaker detection. *arXiv preprint arXiv:2301.08237*, 2023. 8

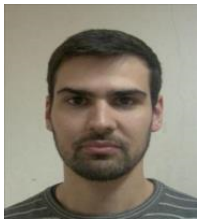
[53] Max Welling and Thomas N Kipf. Semi-supervised classification with graph convolutional networks. In *J. International Conference on Learning Representations (ICLR 2017)*, 2016. 2

[54] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 4

[55] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *BMVC*, 2018. 4

[56] Yuanhang Zhang, Susan Liang, Shuang Yang, Xiao Liu, Zhongqin Wu, Shiguang Shan, and Xilin Chen. Unicon: Unified context network for robust active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3964–3972, 2021. 2

[57] Yuan-Hang Zhang, Jingyun Xiao, Shuang Yang, and Shiguang Shan. Multi-task learning for audio-visual active speaker detection. *The ActivityNet Large-Scale Activity Recognition Challenge*, pages 1–4, 2019. 2



Tiago Roxo (Member, IEEE) obtained a bachelor's degree in Computer Science and Engineering from Universidade da Beira Interior (UBI) in 2019 and is currently pursuing a Ph.D.'s degree, with a FCT (*Fundação para a Ciência e a Tecnologia*) scholarship, in the field of Computer Vision and Artificial Intelligence.

Joana Cabral Costa obtained her bachelor's and master's degree in Computer Science and Engineering from Universidade da Beira Interior (UBI) in 2019 and 2021, respectively. She is currently pursuing a Ph.D.'s degree, with a FCT (*Fundação para a Ciência e a Tecnologia*) scholarship, in the field of Computer Vision and Adversarial Attacks.

Pedro R. M. Inácio was born in Covilhã, Portugal, in 1982. Holds a 5-year B.Sc. degree in Mathematics/Computer Science and a Ph.D. degree in Computer Science and Engineering, obtained from the Universidade da Beira Interior (UBI), Portugal, in 2005 and 2009 respectively. The Ph.D. work was performed in the enterprise environment of Nokia Siemens Networks Portugal S.A., through a Ph.D. grant from the Portuguese Foundation for Science and Technology.

He is a professor of Computer Science at UBI since 2010, where he lectures subjects related with information assurance and security, programming of mobile devices and computer based simulation, to graduate and undergraduate courses, namely to the B.Sc., M.Sc. and Ph.D. programmes in Computer Science and Engineering. He is currently the Head of the Department of Computer Science of UBI. He is an instructor of the UBI Cisco Academy.

He is an IEEE senior member and a researcher of the Instituto de Telecomunicações (IT). His main research topics are information assurance and security, computer based simulation, and network traffic monitoring, analysis and classification. He has about 40 publications in the form of book chapters and papers in international peer-reviewed books, conferences and journals. He frequently reviews papers for IEEE, Springer, Wiley and Elsevier journals. He has been member of the Technical Program Committee of international conferences such as the ACM Symposium on Applied Computing - Track on Networking. He was one of the chairs of WISARC 2016.

Hugo Proença (SM'12), B.Sc. (2001), M.Sc. (2004) and Ph.D. (2007) is an Associate Professor in the Department of Computer Science, University of Beira Interior and has been researching mainly about biometrics and visual-surveillance. He was the coordinating editor of the IEEE Biometrics Council Newsletter and the area editor (ocular biometrics) of the IEEE Biometrics Compendium Journal. He is a member of the Editorial Boards of the Image and Vision Computing, IEEE Access and International Journal of Biometrics. Also, he served as Guest Editor of special issues of the Pattern Recognition Letters, Image and Vision Computing and Signal, Image and Video Processing journals.