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*—This document provides additional details to support the main submission. We begin by describing talking and head bounding box annotations in Sections A and B, respectively. Furthermore, we provide visual examples of Wilder Active Speaker Detection (WASD) categories in Section C, and describe the full list of analyzed features for video selection in Section D.**

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

A. Talking Annotations

We design a custom Graphical User Interface (GUI) for active speaker annotations, as shown in Figure 1. While the video is running, visual sliders are displayed (one for each speaker), automatically filled with red to denote absence of talking. We select a speaker by pressing the corresponding number key (*e.g.*, third speaker is selected with "3" key) and change the speaking label (and slider color) via Ctrl key. To pause, rewind, and forward, we use space, left, and right arrows, respectively. Video time is rewinded and forwarded by 5 seconds with each key press. Prior to GUI launch, we manually set a variable regarding the number of speakers for each video.

AVA-ActiveSpeaker Format Conversion. While performing annotations (either speaking or body/face), we save them on a JavaScript Object Notation (JSON) file (1 file per video), containing all the information used (head/body bounding boxes coordinates, person id, and speaking label), grouped by frame name and person. To convert to AVA-ActiveSpeaker format, we follow the authors guidelines [4], where each line of the annotation file relates to a time frame of a person in a video. To obtain the time frames, we start at time 0 with increments of video duration per video frames. Each line has the entity id (video name with person id), time frame, face bounding box coordinates, and speaking label. For body annotations, we replace face bounding boxes coordinates for body ones. The custom annotations and all AVA-ActiveSpeaker Comma-Separated Value (CSV) files are available at https://github.com/Tiago-Roxo/WASD.

B. Head Bounding Box Annotations

The default approach for head bounding box annotation is based on pose data. Using Alphapose [1], [2], [5], we retrieve the x and y coordinates of the right and left ears, right and left eyes, and nose. The head's central point is calculated using the arithmetic mean of the eye-nose reference point (mean of eyes and nose positions) and ears coordinates. Head bounding boxes are centered in the head's central point, with height and width as a fraction of body silhouette height. This fraction is manually set for each video to ensure adequate head area capture. Figure 2 displays examples of reference points used and head bounding box drawing in different scenarios. In conditions where this approach was not entirely suitable (most Surveillance Settings videos), we annotated manually.

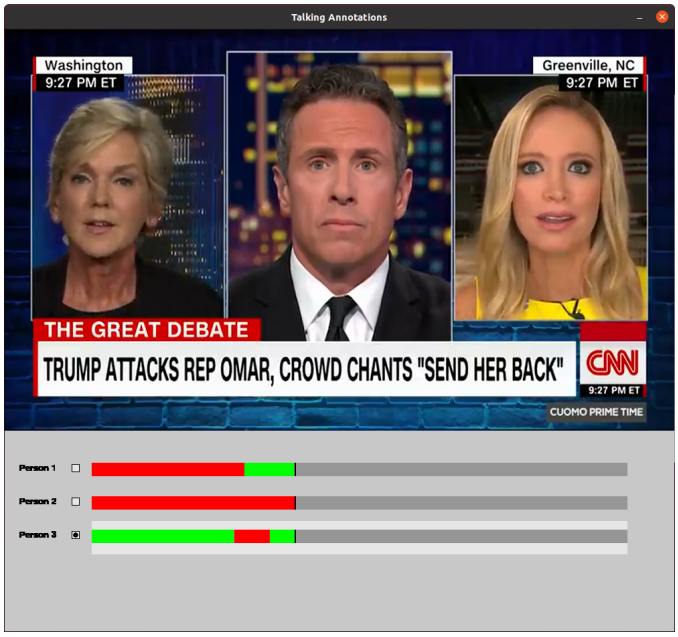


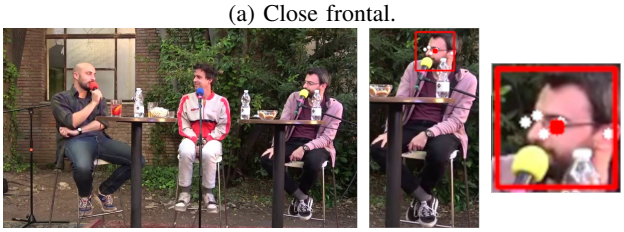
Fig. 1: GUI program used for talking annotations. Three speakers are represented by 3 sliders, below the video. Green color refers to speaking, while red represents silence. Leftmost speaker corresponds to the first slider.

C. WASD Categories

We provide examples of scenarios considered for Wilder Active Speaker Detection (WASD) categories in Figure 3: *Optimal Conditions* mainly consists of interviews or people talking in an alternate manner, with cooperative poses; *Speech Impairment* refers to political debates, heated discussions, and online interviews/debates; *Face Occlusion* contains scenarios where subjects have partial face occlusion from the objects; *Human Voice Noise* relates to subjects reacting to a video, while it plays in the background, contributing to audio impairment; and *Surveillance Settings* are from surveillance interrogations, with variable audio and image quality (*i.e.*, face access and subject cooperation).

Manuscript received XX, 2023; revised XX, 2023.

(a) Close frontal.



(b) Far frontal.

(c) Close sideways.

(d) Far sideways.

Fig. 2: Head bounding box drawing in different scenarios. From left to right, all images contain the original scenario, head bounding box drawing, and zoom in for better visualization. White dots refer to the reference points used for head bounding box drawing, while the red dot is the head's central point. This approach is suitable for various conditions such as close or far frontal poses (a and b), even with facial occlusion, and in side poses at closer or farther positions (c and d, respectively).

D. WASD Features

The complete list of all the considered features, and their admissible values, are the following:

- **Facial Occlusion**: Yes or No;
- **Human Voice as Background Noise**: Yes or No;
- **Speech Overlap**: None-Low or Medium-High;
- **Delayed Speech**: Yes or No;
- **Surveillance Settings**: Yes or No;
- **Body Access**: Low, Medium, or High;
- **Audio Quality**: Low or High;
- **Face Availability**: Guaranteed or Non-Guaranteed;
- **Number of People**: from 2 to 7;
- **Number of White People**: from 0 to 5;
- **Number of Afro People**: from 0 to 4;
- **Number of Asian People**: from 0 to 5;
- **Language**: English, European, or Asian;
- **Number of Females**: from 0 to 4;
- **Number of Males**: from 0 to 5;
- **Frames per Second (FPS)**: from 10 to 30;
- **Image Size**: Variable;
- **Video Location**: Indoor or Outdoor;
- **Body-Image Proportion**: Variable;
- **Head-Body Proportion**: Variable;
- **Speaking Percentage**: Variable;
- **Speaking Overlap**: Variable;
- **Luminosity**: Variable.

All features with predefined admissible values were attributed by human assessment. Although FPS has various admissible values, they mainly range from 24-30. All other features are continuous values, thus not having a strict set of possibilities: Head-Body Proportion refers to the proportion of head area (bounding box) relative to body area, with similar analogy for Body-Image Proportion; Speaking Percentage and Speaking Overlap is the number of frames with talking and simultaneous talking, respectively; and Luminosity is calculated using the red, green, and blue channels to measure the perceived brightness [3]. Regarding Body Access, we consider three values relating to the visible body area and subject proximity to camera. The CSV containing all the information for each WASD video is available at https://github.com/Tiago-Roxo/WASD. Regarding the considered languages, we group them as follows:

- **English**: English (USA and UK);
- **European**: Croatian, Dutch, French, German, Italian, Portuguese, Russian, and Spanish;
- **Asian**: Chinese, Japanese, Korean, Pakistanese, and Vietnamese.

References

[1] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 1

[2] 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. 1

[3] Alexandre Morgand and M. Tamaazousti. Generic and real-time detection of specular reflections in images. *VISAPP 2014 - Proceedings of the 9th International Conference on Computer Vision Theory and Applications*, 1:274–282, 01 2014. 2

[4] 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

[5] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *BMVC*, 2018. 1

(a) Optimal Conditions.

(b) Speech Impairment.

(c) Face Occlusion.

(d) Human Voice Noise.

(e) Surveillance Settings.

Fig. 3: Different examples of the considered scenarios for WASD categories.