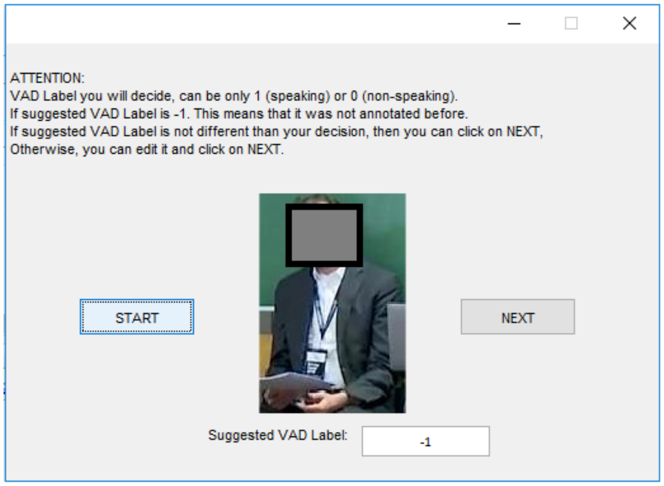
RealVAD: A Real-world Dataset and A Method for Voice Activity Detection by Body Motion Analysis-Supplementary Material

Cigdem Beyan*, *Member, IEEE,* Muhammad Shahid*, Vittorio Murino, *Senior Member, IEEE*



ATTENTION:
VAD Label you will decide, can be only 1 (speaking) or 0 (non-speaking).
If suggested VAD Label is -1. This means that it was not annotated before.
If suggested VAD Label is not different than your decision, then you can click on NEXT,
Otherwise, you can edit it and click on NEXT.

Suggested VAD Label: -1

Fig. 1: The graphical user interface that was used for the second stage of the annotation of the RealVAD dataset. The face of the panelist is hidden only in this figure, i.e., was not hidden during the labeling process.

TABLE I: RealVAD dataset: the number of speaking, not-speaking, discarded bboxes, and the total number of publicly available bboxes having the VAD ground-truth (the sum of second and third columns) for each panelist. IM stands for imbalanced ratio of the publicly available dataset, which is calculated for each person as the total number of speaking bboxes divided by the total number of not-speaking bboxes.

	#speaking	#not-speaking	#discarded	#total	IM
Panelist 1	7796	36829	0	44625	0.21
Panelist 2	8163	29369	7093	37532	0.28
Panelist 3	6401	31597	7093	37998	0.20
Panelist 4	6983	45961	18300	52944	0.15
Panelist 5	8319	45433	17063	53752	0.18
Panelist 6	6665	37960	1	44625	0.18
Panelist 7	7968	24135	22592	32103	0.33
Panelist 8	8981	24298	21962	33279	0.37
Panelist 9	7379	36421	825	43800	0.20

I. Introduction

This supplementary material contains the graphical user interface, which was used during the annotation of our dataset that is called RealVAD. Additionally, more insides regarding the content of our dataset, e.g., the number of speaking and not-speaking person detections as bounding boxes (bboxes) and example bboxes showing freely behaving panelists, occlusions, and background motion, are given.

As the second stage of the VAD labeling, annotators were used the graphical user interface shown in Figure 1. Once the annotator pressed the START button, first 30 consecutive bboxes (equals to one second) belonging to Panelist-1 were displayed. If these bboxes were from the frames moderator was labeled as "speaking" according to the first stage of the annotation (see the main paper for more information), then the suggested label was shown as zero (0) to the annotator as there are no simultaneous speakers. Otherwise, it was shown as minus one (-1). Once the annotator entered the VAD label into the editable text box, if it was one (speaking), then for the other eight panelists, we saved the VAD label as zero, which was retrieved as the suggested label during the corresponding panelists' annotation. While displaying the consecutive bboxes, we did not apply any overlapping. If the annotator agreed with the suggested label, she did not need to enter the label. When annotator pressed the NEXT button, the label written in the editable text box was read and saved as the VAD label of all the 30 bboxes displayed.

In this annotation scheme, the very short pauses, e.g., the one happens while speaker is breathing, might not have been caught. However, since our dataset does not include simultaneous speakers, contains only monologues and does not have fast speech turns like happens in a conversation, we believe that the performed annotation is suitable. Besides, it is worth to recall that between two panelists' speech there is always the speech of the moderator, which was detected in the first stage of the annotation.

In Table I, for each panelist, the number of speaking, not-speaking, discarded bboxes and the total number of publicly available bboxes having VAD ground-truth (the sum of speaking and not-speaking bboxes) are given. Moreover, for each panelist, the class imbalanced ratio (IM), which is calculated as the total number of speaking bboxes divided by the total number of not-speaking bboxes, is also reported. The average numbers of the speaking and not-speaking bboxes in the

* Cigdem Beyan and Muhammad Shahid have equally contributed to this paper. C. Beyan, and M. Shahid are with Pattern Analysis and Computer Vision Research Line, Istituto Italiano di Tecnologia, Genoa 16152, Italy (e-mail: cigdem.beyan@iit.it, shahid.muhammad@iit.it). M. Shahid is also with Dipartimento di Ingegneria Navale, Elettrica, Elettronica e delle Telecomunicazioni, University of Genova, Genoa 16145, Italy. V. Murino is with the Department of Computer Science, University of Verona, Verona 37129, Italy, Huawei Technologies LtD., Ireland Research Center, Dublin, Ireland, and Pattern Analysis and Computer Vision Research Line, Istituto Italiano di Tecnologia, Genoa 16152, Italy, (e-mail: vittorio.murino@iit.it).

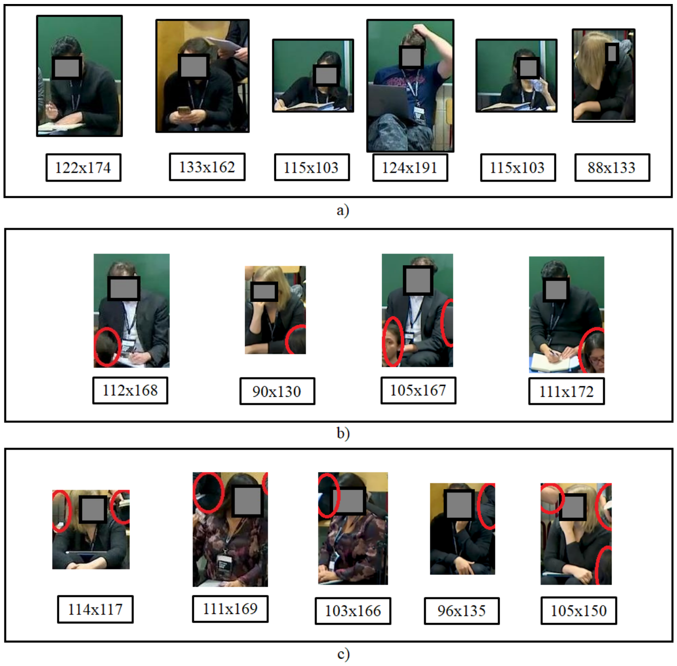


Fig. 2: RealVAD dataset: Some example bboxes showing *a)* panelists are freely moving, i.e., doing various actions resulting in different postures, *b)* occlusions, and *c)* background motion. The faces of panelists were hidden with black rectangles filled with grey color. The red ellipses in b) and c) show the occlusions between a panelist and another panelist's head, laptop, etc., and the background motion occurring when a panelist in the back-row moves, respectively. The numbers show the resolution (width×height) of the bboxes.

publicly available dataset are 7628.3 and 35495.8, respectively. The average of IMs is 0.23.

Overall, the discarded bounding boxes are due to *i)* the disagreement between annotators or *ii)* very low IM between the speaking and not-speaking bboxes (implying fewer speaking bboxes than not-speaking bboxes). In that table, the given numbers correspond to *(ii)*, which was in a way performed arbitrarily but considering two aspects. The first one is the IM per panelist. Having so many not-speaking bboxes would not contribute to the behavior diversity much, and conversely, introducing redundancy. As a consequence of this, training a model was taking very long time without contributing the learning (in our case this refers to ResNet50 fine-tuning). Still, we tried to have variable numbers of IM, as this can be challenging for some of the VAD methods and valuable to test. In the final dataset, there are three panelists having IM<0.20, four panelists having 0.20≤IM<0.30 and two panelists having 0.30≤IM<0.40. These thresholds were arbitrarily chosen. The second considered aspect was not discarding the intermediate bounding boxes as this would create discrete and shorter segments. With the same motivation, the removal was performed first on the very short segments, which were appeared due to discarding the bboxes having annotator disagreement.

Some example bboxes illustrating: *a)* freely moving panelists (i.e., doing various actions resulting in different postures), *b)* occlusions, and *c)* background motion, are given in Figure 2. The resolutions of these bboxes are also reported. As seen, the resolutions of bboxes can be various even for the same panelist. Each of these situations can individually make the VAD task challenging.