S-VVAD: Visual Voice Activity Detection by Motion Segmentation

Muhammad Shahid[1,2], Cigdem Beyan[1], Vittorio Murino[1,3,4]
{shahid.muhammad, cigdem.beyan, vittorio.murino}@iit.it
[1]Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, Italy
[2] DITEN, Università degli Studi di Genova, Italy
[3]Department of Computer Science, Università di Verona, Italy
[4]Huawei Technologies Ltd., Ireland Research Center, Ireland

Abstract

We address the challenging Voice Activity Detection (VAD) problem, which determines "Who is Speaking and When?" in audiovisual recordings. The typical audio-based VAD systems can be ineffective in the presence of ambient noise or noise variations. Moreover, due to technical or privacy reasons, audio might not be always available. In such cases, the use of video modality to perform VAD is desirable. Almost all existing visual VAD methods rely on body part detection, e.g., face, lips, or hands. In contrast, we propose a novel visual VAD method operating directly on the entire video frame, without the explicit need of detecting a person or his/her body parts. Our method, named S-VVAD, learns body motion cues associated with speech activity within a weakly supervised segmentation framework. Therefore, it not only detects the speakers/not-speakers but simultaneously localizes the image positions of them. It is an end-to-end pipeline, person-independent and it does not require any prior knowledge nor pre-processing. S-VVAD performs well in various challenging conditions and demonstrates the state-of-the-art results on multiple datasets. Moreover, the better generalization capability of S-VVAD is confirmed for cross-dataset and person-independent scenarios.

1. Introduction

Voice Activity Detection (VAD) answers the question "Who is Speaking and When?" in audiovisual recordings. VAD is a key process for a number of technologies, e.g., multiparty dialog problem in human-human [16] and human-robot interaction systems [18], social behavior analysis [4], automatic speech recognition [13], speech enhancement [34], and emotion recognition [37]. Since VAD is applied at the initial stage of such technologies, its accuracy and robustness are essential. Despite its crucial role in such application domains, VAD remains a challenging problem due to wide diversity of speakers, the possible presence of multiple persons and simultaneous speakers as well as the inconvenient position of persons badly located with respect to the microphone/camera, which may affect the input data quality [15].

The traditional way to perform VAD is based on audio signal processing. Audio-based VAD systems are challenged by the presence of ambient noise, and they are sensitive to noise variations that are very common in real-world conditions [29]. Moreover, they suffer from complications occurring when there is overlapping speech [17]. Recently, multimodal VAD approaches become attractive due to their more accurate performances, e.g., [8, 25, 33, 2]. However, multimodal VAD might not be applicable when the audio data is not available, e.g., due to technical, ethical or legal issues. In such cases, the use of information from visual modality only, so-called "visual VAD" is very desirable.

Visual VAD (VVAD) literature is limited as compared to audio-based solutions. Almost all VVAD approaches require body part detection, which brings extra computational cost and the success of VAD is dependent to the success of the body part detectors. Mostly, face detection [32, 17, 25, 2] or specifically lips detection has been used [31, 19, 14, 8]. Lately, it is shown that during speech there are important visual cues beyond face and lips movements, e.g., head activity [11], hand motion [16, 11], gaze [16], and upper body motion [7, 27, 28, 5], which are effective cues to be considered.

Unfortunately, to date, VVAD methods have usually been evaluated on datasets either having a single-person frontal view [16, 32] or supplying the person/face extracted [7, 25, 2, 27, 28, 5]. Nevertheless, VAD methods should cope with more difficult scenarios and situations in practical applications, such as multiple speakers' turns, variable person/face pose, camera position and environmental conditions [17].

In this paper, we present a novel VVAD methodology

(Section 3), named Visual Voice Activity Detection by motion Segmentation – S-VVAD. Given that the importance of modeling body motion during speech has been acknowledged in several studies [16, 17, 7, 28, 28, 29, 5], we decided to examine this information more thoroughly for targeting VAD. This aspect is of clear relevance in the current global pandemic emergency as the lips-based methods (the most frequently used body part for realizing VAD) would completely fail in case of wearing a face mask, while body motion-based VAD can still perform well.

S-VVAD utilizes the "holistic" body motion of a person (without relying on a specific body part) represented in terms of dynamic images [6]. The body motion cues associated with speech activity (speaking/not-speaking) is learned in conjunction with a weakly supervised segmentation scheme. Therefore, S-VVAD is not only able to determine whether someone is speaking or not, but also localizes that person simultaneously.

S-VVAD processes without any usage of prior information, e.g., training samples of faces [32], having information identifying the mouth as the lower half of the video frames [21], knowledge of the number of persons present [30] or the structure of the conversation (e.g., small group meetings [16]). Also, it does not require the face onscreen [25, 29, 2] nor it does need synchronized video and audio tracks [8, 25, 2], or access to the subtitles [29]. Moreover, even though it is proved that person-dependent models perform substantially better than person-independent models [17], S-VVAD is person-independent and generic, meaning that it can be applied to any new person without re-training. Finally, it is an end-to-end method, in which visual features are directly learned from the data, it works on the entire video frame without applying additional person detection at test time (unlike all methods except [29]). All these features are very important and have never been addressed altogether by any of the existing VAD approach.

S-VVAD applied to real-world datasets having single/multiple persons in a video frame shows performance not only better than the state-of-the-art (SOA) VVAD methods, but also better than the SOA multimodal VAD approaches. Cross-dataset experiments show that our method has an effective potential to be applied to different real-world situations, while being trained on other settings.

The contributions and highlights of S-VVAD can be summarized as follows:

- This is the first attempt that VVAD is performed by body motion cues learning with weakly supervised segmentation. As a result, S-VVAD is able to determine speech activity of a person while simultaneously localizing her/him. It should be noted that so far such localization have been performed only by using speaker diarization techniques [3], which are based on audio processing. Whereas herein we perform this in visual domain.
- S-VVAD demonstrates effective results when tested on various conditions: single/multiple subjects in the same video frame, simultaneous speakers/silence, occlusions, background motion and so forth.
- S-VVAD is practical: it does not require any prior knowledge, does not have any constraints on the number of persons as well as it is person-independent (can be applied to any new person without re-training). Moreover, it is end-to-end, processing on the whole video frame in the test time, not relying on accurate detection of body parts (lips, hands, etc.), and learns the features directly from the data.

2. Related Work

Visual Voice Activity Detection. VVAD has been performed using facial cues (e.g., facial landmarks movement, lips motion), body motion cues (e.g., hand gestures, head movement, upper body motion), or a combination of both. As an earlier work, Liu and Wang [23] utilize handcrafted features describing lips region. Head and lips motion are combined in [17], showing that lips motion performs the best. For human-machine interaction, [14] indicated that head motion is a significant cue while its fusion with lips motion performs better. More recently, face features extracted from a CNN, are modelled with an LSTM to predict the speech status of a face [32]. These methods [23, 17, 14, 32] might be ineffective when face/lips detection fail, e.g., when the speaker presents a profile view to the camera, she/he is far away from the camera, her/his face is occluded or the camera resolution is low.

On the other hand, Hung et al. [16] use hand motion and gaze (implying head orientation) to perform VAD in small group meetings. Hand motion is examined assuming that the speaker is the one who moves the most. Gaze is investigated by claiming that the majority of the subjects should be looking towards the speaker. Results demonstrate that gaze can be a good indicator to detect the speaker [16]. However, gaze might not be feasible for the scenarios, which people sit in a single row, as it is rare that they face each other. Additionally, accurate gaze estimation is very challenging and its implementation in-the-wild is still an ongoing research topic. Later on, for the same dataset, it was shown that motion history images extracted from upper body performs as good as gaze [12].

For the first time, real-world datasets (not a role-play) were used to test a VVAD approach in [7]. Audio is used for cross-modal supervision of the training of a personalized VVAD, which is based on improved trajectory features of upper body [7]. Using the same dataset, several body motion descriptors (e.g., optical flow images, dynamic images [6]) are compared for person-independent VVAD in [27]. The results show that RGB-based dynamic image represen-

tation of upper body performs the best [27]. Subsequently, in [28, 5], RGB-based dynamic images are combined with an unsupervised domain adaptation technique, demonstrating improved results compared to [7, 27, 8].

As can be noticed, the majority of the methods are based on body part detection [16, 23, 17, 14, 32]. Furthermore, some of them [23, 16, 12] were tested on images having one person, while others [7, 27, 28, 5] used highly accurate, manually extracted person/face detections in training and test. Recently, the first VVAD approach that does not require additional person/face detection was presented in [29]. This method operates on the entire video frame: in training a video of an individual person is divided into short segments, which are modelled with unconnected 3D-CNNs and then combined with an LSTM [29]. Similarly, our method does not apply separate/additional person detection but, unlike [29], we use RGB-based dynamic images, summarizing the motion and related appearance of a video segment. Also, we learn localizing the VAD-related motion cues within a weakly supervised segmentation framework.

3. The S-VVAD

First, body motion cues representing a speech activity are learned by fine-tuning a ResNet50 (Section 3.1, Figure 1C1). Then, that model is used to obtain class activation maps (CAMs), which are further used for the training of a Fully Convolutional Network (FCN), performing segmentation as speaking, not-speaking and background (Section 3.2, Figure 1C2). In test, the trained FCN is used to localize subjects and their speech activity in an input video segment (Section 3.3, Figure 2). These are described in detail as follows and illustrated in Figures 1 and 2. Code is publicly available at[1].

3.1. ResNet50 Fine-tuning with Body Motion Cues

Recent works [27, 28, 5] have showed that body motion represented in terms of dynamic images (DIs) [6] can be very effective for VVAD (see [27] for VVAD performance comparisons among different motion representation methods; optical flow images (OFI), RGB-based DI, OFI-based DI and, etc.). Moreover, training a deep model with DIs as compared to RGB frames is less computational complex as one DI is built from multiple frames. We have adapted that representation and obtained one DI from 10 consecutive RGB frames. The resulting image summarizes the short-term spatio-temporal content, i.e., the body motion of the subjects (also background motion if any) and the appearance associated with these. A DI is, then cropped into sub-dynamic images (sub-DIs), each contains a single subject. In our implementation, we do not apply image cropping in the way that it tightly surrounds a subject but instead we include the background as well (i.e., the length of vertical axes of the sub-DI and the corresponding DI are the same).

Later, we fine-tune ResNet50 (pre-trained on ImageNet) with multiple sub-DIs with speaking/not-speaking VAD labels, as follows. The input sub-DIs are either *i)* resized to 256×256, which is called single-scale setting or *ii)* each of them is first resized to K×K when K is a randomly generated number between 256 to 320, and then a random 256×256 chunk of obtained image is used as the input. The latter strategy is called as multi-scale setting, which has not been applied in [27, 28, 5]; however, it improves the VAD performance of S-VVAD. Unlike [27, 28, 5], we train the convolutional layers shown as Block 3, 4, 5 in Fig. 1 and the Softmax layer. The weights of blocks are randomly initialized before training. Training is performed with a decay learning rate e.g., starting from 5×10^{-6}, using cross entropy loss function, and Adam optimizer while in each batch 128 randomly selected sub-DIs (64 speaking and 64 not-speaking) are used. Data augmentation is also applied such that some randomly selected sub-DIs are horizontally flipped and/or a 64×64 randomly selected patch is replaced with the mean value of the images. 10% of the training data is randomly selected as the validation data and when the loss on validation data increases, training is stopped.

3.2. CAM-based Mask Generation & FCN Training

Class activation maps (CAMs) have been actively exploited in many computer vision tasks to explain what a DNN architecture learns [35, 36, 10]. These maps provide a viable option for trained models to pinpoint the region of interest of an image. We adapted Grad-CAM [26], which needs the target class gradient information reaching to the final convolution layer and the convolution feature maps of trained model.

Given that 10 consecutive RGB frames are used to constitute one DI; X containing t number of participants with labels $C_l = [C_1, ..., C_t]$ when $C_i \in [0, 1]$ (0:not-speaking, 1:speaking), first, we apply the trained ResNet50 (Section 3.1) to obtain the last layer convolution feature maps $\Upsilon^J = \psi(X, \theta)\epsilon R^{J\times U\times V}$; ψ is a function parameterized with θ, U, V represent the size and J is the number of feature maps. In case of ResNet50, Υ is 32 times downsampled with respect to the size of DI. Afterwards, the weights are computed using the gradient of speaking class score y^S (before the softmax) and not speaking class score y^{NS} (before the softmax) corresponding to the convolution feature maps given in Eq. 1. For each class, a CAM ($\lambda^C_{(u,v)}$), is computed using the convolution feature maps and the computed weights as in Eq. 2, when Z is equal to $U \times V$.

$$W_j^S = \frac{1}{Z}\sum_u\sum_v \frac{\partial y^S}{\partial \Upsilon^j_{uv}}, \qquad W_j^{NS} = \frac{1}{Z}\sum_u\sum_v \frac{\partial y^{NS}}{\partial \Upsilon^j_{uv}} \tag{1}$$

$$\lambda^S_{(u,v)} = \sum_j W_j^S \Upsilon^j_{uv}, \qquad \lambda^{NS}_{(u,v)} = \sum_j W_j^{NS} \Upsilon^j_{uv} \tag{2}$$

[1] github.com/IIT-PAVIS/S-VVAD

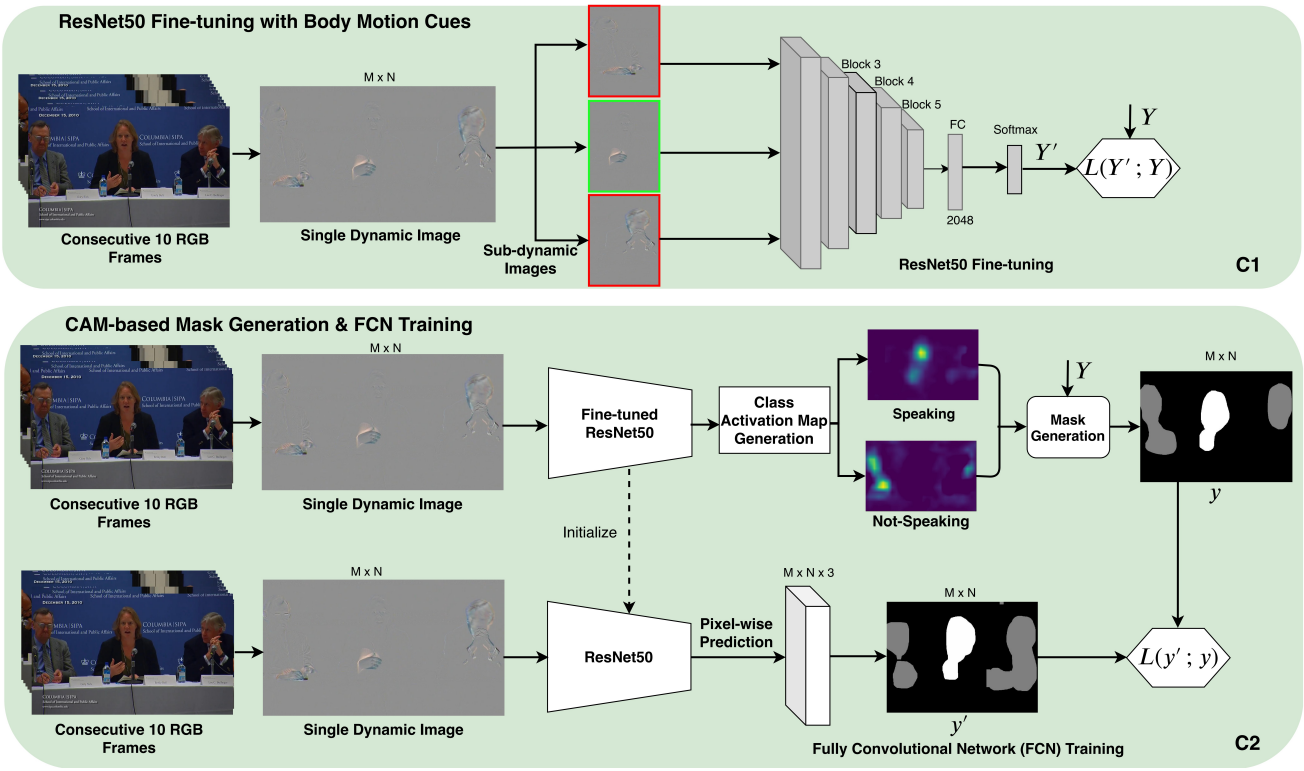

Figure 1. S-VVAD has successive two components: C1) ResNet50 fine-tuning with body motion cues: a ResNet50 is fine-tuned with sub-dynamic images (sub-DIs) having labels either speaking (green bounding box) or not-speaking (red bounding box). Each sub-DI includes spatio-temporal information of one subject. C2) CAMs-based mask generation and FCN training: CAMs; one for speaking, other for not-speaking are obtained when a dynamic image (DI; M×N) is the input to the ResNet50 trained in C1. A mask (M×N), shown as y, is formed using these CAMs and the ground-truth (Y). An FCN composed of ResNet50 pre-trained in C1, is trained with DIs (M×N). For each DI, FCN makes a pixel-wise prediction (M×N×3) with channels: background, speaking or not-speaking. The pixel-wise cross entropy loss (L) of FCN is calculated between y and pixel-wise prediction.

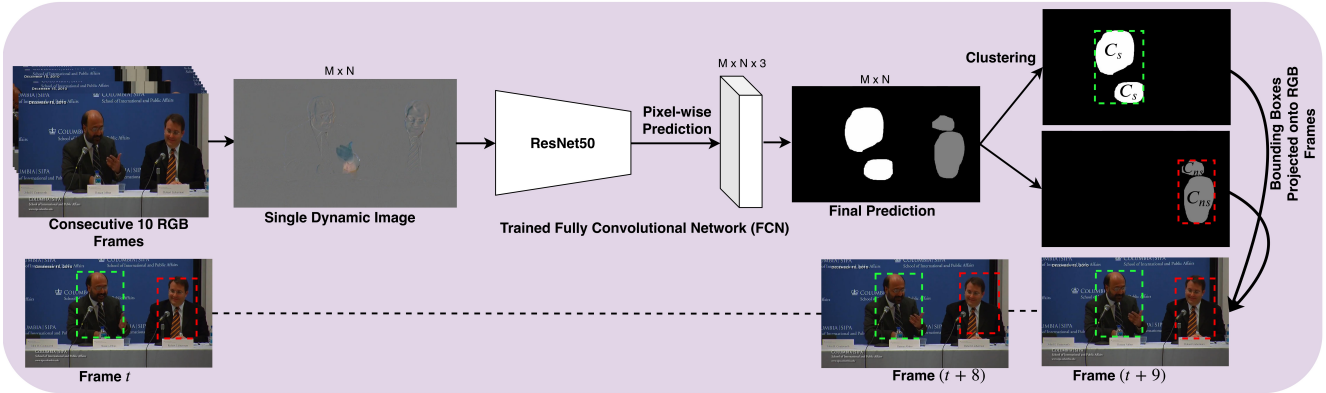

Figure 2. At test time, a single dynamic image (M×N), constructed from 10 consecutive RGB frames, is first segmented using the trained FCN. Segmentation results in an image called as final prediction (M×N) containing pixels predicted as: background (shown in black), speaking (shown in white) and not-speaking (shown in grey). The location of speaking and not-speaking pixels are clustered individually. Pixels belonging to the same cluster are merged into a bounding box form. Then, the location of these bounding boxes are projected onto 10 RGB frames. Green and red dashed bounding boxes indicate the speaking and not-speaking persons detected, respectively.

Then, CAMs ($\lambda^{C}_{(u,v)}$) are upsampled to the size of DI using bi-cubic interpolation. Given the image obtained from $\lambda^{S}_{(u,v)}$, we keep the pixels, which intersect with the pixels belonging to the speaking persons in the ground-truth, as they are, and set all others to zero. The same procedure is applied to the image obtained from $\lambda^{NS}_{(u,v)}$ as well. This is

so-called using weak annotations [24]. The two images obtained are then thresholded separately using OTSU method [22], resulting in two binary images. To obtain a mask, these binary images are merged such that speaking pixels are multiplied by 255 and not-speaking pixels are multiplied by 127. A mask can have pixels having up to three semantics: speaking, not-speaking and background. Unlike some studies [36, 10], we do not re-train the used DNN recursively to improve CAMs. Instead, it is effective to use the ground-truth and apply independent thresholding to create the masks.

Finally, the trained ResNet50 (Section 3.1) is integrated to FCN-8 network architecture [20]. The final fully connected layer of ResNet50 is removed. Starting from Block 5, an upsampling is applied to reach the size of Block 4. The result is summed up with Block 4 and then upsampled to reach the size of Block 3. That result is then summed up with Block 3. In that way, a multi-channel feature map having the same size of the input image is obtained. The final layer of FCN-8 is a 1×1 convolutional layer, producing three channels having the meanings: speaking, not-speaking and background (called pixel-wise prediction in Fig. 1).

This described network is trained with an end-to-end manner when a DI is the input, the corresponding generated mask is the ground-truth and the pixel-wise cross entropy loss is calculated between the mask and pixel-wise prediction. As masks are generated from weak annotations, the performed segmentation is weakly supervised. Also, this is the first time that DIs are being used within a segmentation problem. The Adam optimizer with learning rate of 10^{-5} is applied until the loss in validation data (see Section 3.1 for description) increases. For the training of FCN, two settings are applied. In the first setting, ResNet50 is freezed such that its weights are not updated through FCN training. Whereas in the second setting, Block 3, 4 and 5 of ResNet50 are updated. The latter setting results in improved VAD performances.

Examples of dynamic images, class activation maps (CAMs) and masks, which are obtained during the training of S-VVAD, are given in supplementary material.

3.3. S-VVAD in Test Time

At test time, first, a single DI is constructed from 10 consecutive RGB frames. Using the trained FCN, that DI is segmented such that pixel-wise prediction having three channels: background, speaking, and not-speaking, is obtained. These predictions are merged into a single image as applied for mask generation in Section 3.2 and called as final prediction (see Figure 2). The size of final prediction image is the same with the input DI (as well as the RGB frames), and it contains pixels having one of the three semantics: background, speaking or not-speaking. The location of speaking pixels and not-speaking pixels are clustered individually using Affinity Propogation (AP) [9] algorithm. We preferred using AP as it does not require the number of clusters. For each cluster (recall that each cluster is either speaking or not-speaking), we find a bounding box (bbox), which tightly surrounds it. In an optimum scenario, a bbox should correspond to a person in the scene and VAD label of the cluster indicates whether she/he is speaking or not. The location of bboxs are projected onto the 10 RGB frames with the VAD labels.

To evaluate S-VVAD, we calculate the intersection-over-union (IoU) score between projected bboxes and ground-truth bboxes (showing the location of persons). Only the bboxes having IoU > 0.5 is used to further compare the detected VAD label with the ground-truth. Thus, the bboxes that does not obey the IoU rule are directly considered as misclassified VAD (i.e., false negative or false positive).

Figure 3. Example video frames from datasets used. All datasets are from a real-world panel discussion. (a) Columbia dataset [7]: supplies the image position of each panelist's upper body. (b) Modified Columbia dataset: there are two/three sitting panelists in a video frame freely moving, camera motion exist in some video frames. (c) RealVAD dataset [5]: Nine panelists in a video frame. They are freely moving. Occlusions and background motion exist. The distances between panelists and camera are diverse. More images can be seen in supplementary material Figure 5.

4. Datasets

S-VVAD was tested using all publicly available VAD datasets constructed from real-world situations, which supply the ground-truth of body location of persons (Fig. 3). In other words, the datasets that *i)* were captured in lab environments, having simple and unrealistic scenarios as well as *ii)* the ones providing only the location of face or lips were not included. The lattermost is particularly important to perform fair evaluation and comparison.

4.1. Columbia Dataset

Columbia dataset [7] was constituted from a YouTube video of a panel discussion. The annotated data is 35 minutes, containing image location of the panelists (bounding box ground-truth), and VAD ground-truth (speaking/not-speaking). This corresponds to in total five panelists (Bell, Bollinger, Lieberman, Long, Sick). In some of the frames camera motion exist.

As the evaluation protocol F1-score and leave-one-panelist-out cross validation (at each fold of cross validation, training set contains the data belonging to four panelists and test set includes the data belonging to one remaining panelist) were used. In that setting, S-VVAD was trained and tested on images having one panelist. We created one DI from 10 consecutive RGB frames. All these settings are in line with SOA methods given in Table 1.

4.2. Modified Columbia Dataset

In order to evaluate S-VVAD in real-world scenarios having multiple persons in a video frame, we have modified the Columbia dataset as follows. First, we extracted the video frames having VAD ground-truth. This resulted in images containing either two or three panelists, each has an individual VAD label. Then, we assigned these images into cross validation folds such that the panelists in a training fold and the corresponding test fold do not overlap. Such a setting decreased the number of frames in the training folds as compared to the setting in Section 4.1. The total number of frames and panelists differ at each cross validation folds, i.e., some training folds have more data than others. We obtained six folds, named as groups, and the cross validation applied is referred as leave-one-group-out. This data split is available here[2]. The evaluations are still performed in person-level as the task is detecting whether each person in a video frame is speaking or not, while each test image has either two or three VAD labels depending on the number of panelists it contains.

4.3. RealVAD Dataset

RealVAD dataset is constructed from a YouTube video composed of a panel discussion lasting for approx. 83 minutes [5].There is only one static camera capturing all panelists, the moderator and audiences. The VAD annotations belong to 9 panelists who are sitting in two-rows. Panelists are not looking at the camera but instead they can be gazing audience, other panelists, their laptop, the moderator or anywhere in the room while speaking or not-speaking. Thus, they were captured not only from frontal-view but also from side-view varying based on their instant posture and head orientation. It is possible to observe panelists doing various spontaneous actions (e.g., drinking water, checking their cell phone, using their laptop, etc.), resulting in different postures. Their body parts are sometimes partially occluded by their/other's body part or belongings (e.g., laptop). The distance between camera and each panelist is also varied. The background of the front and back row panelists are also different. For the panelists sitting in the front row, there is sometimes background motion occurring when the person(s) behind them moves. There are also natural changes of illumination and shadow rising on the wall behind the panelists in the back row.

We follow the same evaluation protocol with the SOA (F1-score and leave-one-panelist-out cross validation) while we created one DI from 10 consecutive RGB frames that is in line with SOA.

[2] github.com/IIT-PAVIS/S-VVAD

5. Experimental Analysis & Results

The experimental analysis include: *i)* comparisons with SOA VVAD and multimodal VAD approaches, *ii)* an ablation study, *iii)* cross-dataset experiments and *iv)* qualitative assessment.

5.1. Comparisons with the State-of-the-Art

Table 1 compares the performance of different methods using *Columbia dataset* [7]. Studies [7, 27, 28, 5] perform body motion-based VVAD. Chung and Zisserman [8], Roth et al. [25] and Afouras et al. [1] apply multimodal VAD. Studies [28, 5] include unsupervised domain adaptation. The results of [28, 5] given in Table 1 corresponds to the setting that whole testing data was input to the unsupervised domain adaptation component. Chung and Zisserman [8] models audio and lips motion jointly (so-called SyncNet). Roth et al. [25] uses three architectures: a) visual embedding network (CNN) in which face images are the inputs, b) audio embedding network (CNN) and c) prediction network (either fully connected layers followed by softmax or a Gated Recurrent Units followed by fully connected layer and softmax). Afouras et al. [1] presents a model using audio-visual attention to localize sound sources and aggregate the extracted information over time via optical flow.

As seen in Table 1, our method (S-VVAD) outperforms all the SOA approaches, on average. Importantly, it has the lowest standard deviation (STD) while performing the best. This shows that S-VVAD is able to generalize well, resulting similar performance independent to test person.

The performance of S-VVAD is a very important achievement given that *i)* it analyzes only one modality: the body motion, thus, does not require synchronization of video and audio (opposite to [8, 25]), *ii)* it operates on whole body of a person without requiring body part detection (opposite to [8, 25, 1]), *iii)* at test time, it does not apply an additional person detection algorithm, instead works on entire video frame (opposite to all SOA), *iv)* during training, it does not rely on bounding boxes tightly surround a person

Table 1. F1-scores (%) on the Columbia dataset. FT, and NA stand for fine-tuning (see the corresponding reference for details) and not-available, respectively. All results were obtained by processing with 10-frame window as in [25]. The best result of each column is emphasized in bold-face.

Method	Bell	Bollinger	Lieberman	Long	Sick	Avg.	Std.
[7]	82.90	65.80	73.60	86.90	81.80	78.20	8.00
[27] (Softmax)	86.07	93.30	91.88	73.62	86.34	86.24	8.00
[27] (SVM)	86.34	93.78	92.34	76.09	86.25	86.96	7.00
[8]	**93.70**	83.40	86.80	**97.70**	86.10	89.54	6.00
[28, 5] (FT1)	87.28	96.35	92.15	83.03	87.21	89.20	5.00
[5] (FT2)	91.92	**98.90**	**94.05**	89.07	92.84	93.36	4.00
[25]	NA	NA	NA	NA	NA	92.80	NA
[1]	92.60	82.40	88.70	94.40	**95.90**	90.80	5.41
S-VVAD	92.36	97.23	92.27	95.50	92.48	**93.97**	2.00

(opposite to [7, 8, 25]), *v)* it is an end-to-end detector and does not require applying domain adaptation to generalize well (opposite to [28, 5]).

Using *Modified Columbia* dataset, the performance of S-VVAD is compared with [5]. Beyan et al. [5] is the best performing VVAD method out of all SOA in Table 1, thus we included it to the comparisons. We also reported the results of random guess, which was computed by randomly generating VAD labels and then calculating the F1-score based on them, for 1000 individual times. The reported score is the average of these 1000 scores.

Modified Columbia is more challenging than Columbia dataset as it includes fewer training data at each fold. Additionally, having multiple persons in a frame introduces extra challenges for S-VVAD. Instead, for [5], bounding box ground-truth (implying perfect person detection) was used in training and test. As seen in Table 2, S-VVAD performed significantly (p-value < 0.05) better than [5] on average as well as performing the best for each group (G1-G6).

Table 2. F1-scores (%) on Modified Columbia dataset. G1-G6 refer to groups (see text). The best result of each row is emphasized in bold-face.

Method	Random Guess	[5]	S-VVAD
G1	42.03	72.09	**86.12**
G2	50.00	72.49	**75.10**
G3	40.00	70.00	**83.96**
G4	39.99	87.49	**87.67**
G5	50.02	83.88	**94.25**
G6	40.02	96.46	**96.73**
Avg.	43.68	80.40	**87.30**
Std.	5.00	11.00	8.00

5.2. Ablation Study

We carried out an ablation study using Modified Columbia dataset to evaluate the benefit of the components of S-VVAD. Corresponding results are given in Table 3.

The ablation study shows that using multi-scale setting in training as compared to single-scale setting (Section 3.1) generally improves the performance of the first component of S-VVAD (C1), resulting in better VAD score on average. FCN was trained with two settings: by freezing ResNet50 or by updating Blocks of ResNet50 (Section 3.2). The results show that the further (S-VVAD) performs better than the former (C1 with multi-scale setting + C2 with freezed ResNet50). Overall, it can be seen that each component has a positive contribution.

5.3. Cross-Dataset Analysis

The cross-dataset analysis performed is: training S-VVAD with the whole Modified Columbia dataset (i.e., the images are composed of two or three persons) and testing the trained S-VVAD on the entire video frames of RealVAD dataset [5]. The corresponding results are given in Table 4, which are compared with the baseline results. The baselines include the method of Beyan et al. [5] that is *i)* trained and tested on RealVAD dataset (so-called the same-dataset analysis) while both training and testing images include a single person at a time and *ii)* trained on Columbia dataset (i.e., the images composed of single person) and tested on RealVAD dataset's images composed of single person. These experimental settings are illustrated in supplementary material in addition to the localization results of S-VVAD for the aforementioned cross-dataset analysis.

S-VVAD performs slightly (52.55%) better than [5] (51.52%) on average, in cross-dataset setting. This is an important improvement particularly, given that *i)* the domain gap between train and test images in our cross-dataset setting (2-3 persons in training, crowd in test) is bigger than the cross-dataset setting in [5] (single person both in training and test) and *ii)* we do not apply any domain adaptation techniques while [5] is based on unsupervised domain adaptation. On the other hand, 53.04% F1-score can be seen as the upper bound as being obtained by the same-dataset analysis. There is no statistically significant performance

Table 3. Ablation study performed on Modified Columbia dataset. G1-G6 refer to groups (see text). See C1 and C2 in Fig. 1. S-VVAD can be written as "C1 with (w/) multi-scale setting + C2". The best result of each column is emphasized in bold-face.

Method	G1	G2	G3	G4	G5	G6	Avg.	Std.
C1 w/ single-scale setting	60.20	68.77	66.81	86.88	78.69	92.72	75.68	13.00
C1 w/ multi-scale setting	70.58	70.78	69.15	82.35	82.14	91.76	77.79	9.00
C1 w/ multi-scale setting + C2 w/ freezed ResNet50	80.51	74.52	81.11	86.88	89.16	94.61	84.47	7.00
S-VVAD (C1 w/ multi-scale setting + C2)	**86.12**	**75.10**	**83.96**	**87.67**	**94.25**	**96.73**	**87.30**	8.00

Table 4. Cross-Dataset Analyses: the test set is RealVAD dataset [5], training set is given in bracket for each method. The best result of each column is emphasized in bold-face.

Method	P1	P2	P3	P4	P5	P6	P7	P8	P9	Avg.	Std.
[5] (Trained w/ RealVAD)	51.63	53.49	42.92	**51.70**	**44.40**	50.48	**58.73**	67.94	55.75	**53.04**	7.50
[5] (Train w/ Columbia)	53.56	51.08	41.09	50.22	37.29	50.32	56.74	53.58	**69.79**	51.52	9.25
S-VVAD (Train w/ Modified Columbia)	**58.33**	**59.28**	**47.97**	44.76	37.28	**57.38**	55.60	**71.34**	41.05	52.55	10.69

difference between our method (52.55%) and the result of the same-set analysis (53.04%).

5.4. Qualitative Assessment

Example qualitative VAD and localization results of S-VVAD are given in Figure 4. These results were obtained when our method was applied to Modified Columbia dataset. More results including S-VVAD applied to other datasets can be found in supplementary material.

In Figure 4, the localization results imposed on the middle RGB video frame (left) -recall that one dynamic image is obtained from 10 consecutive RGB video frames- and then the corresponding bounding boxes (right), if and only if they are classified correctly, are shown. S-VVAD is able to differentiate the body motion of the speakers and non-speakers, which can be observed from the localization results. Some activities resulting in body motion are changing the head and/or body pose, opening a bottle, drinking water. They have been correctly differentiated from the body motion due to speaking. On the other hand, there are some cases, where the localization is performed correctly, i.e., S-VVAD is able to detect the body motion associated with speech activity correctly, however, the predicted bounding boxes are not plotted. This is because the predicted bounding boxes do not supply the IoU rule applied.

Figure 4. Example localization results (red for not-speaking, green for speaking) imposed on the middle RGB video frame (left). The corresponding predicted bounding boxes (red for not-speaking, green for speaking), which are correctly detected (right).

6. Conclusions & Future Work

We have demonstrated a novel visual voice activity detector, named S-VVAD, which learns body motion cues related to speech activity within a weakly supervised segmentation. S-VVAD works on a single modality, is end-to-end, models the body motion of a person without relying on body part detectors. At test time, it operates on the whole image, i.e., it does not apply additional person detectors to localize a person. It is very practical as it works without any prior information, does not have constraints on the number of persons present in a scene or on the structure of the interaction. It is a generic, person-independent approach.

S-VVAD was validated on various challenging conditions and demonstrated improved results on multiple datasets. It is able to generalize well, thus, can be applied to different datasets while being trained on others. S-VVAD will be extended to process on distributed camera systems.

References

[1] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman. Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision*, 2020.

[2] J. L. Alcazar, F. Caba, L. Mai, F. Perazzi, Joon Y. Lee, P. Arbelaez, and B. Ghanem. Active speakers in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[3] O. Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. Speaker diarization: A review of recent research. *IEEE Trans. on Audio, Speech, and Language Processing*, 20(2):356–370, 2012.

[4] C. Beyan, V.-M. Katsageorgiou, and V. Murino. A sequential data analysis approach to detect emergent leaders in small groups. *IEEE Trans. on Multimedia*, 2019.

[5] C. Beyan, M. Shahid, and V. Murino. RealVAD: A real-world dataset and a method for voice activity detection by body motion analysis. *IEEE Transactions on Multimedia*, Early Access.

[6] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould. Dynamic image networks for action recognition. In *CVPR*, 2016.

[7] P. Chakravarty and T. Tuytelaars. Cross-modal supervision for learning active speaker detection in video. In *ECCV*, pages 285–301, 2016.

[8] J. S. Chung and A. Zisserman. Learning to lip read words by watching videos. *CVIU*, 173:76–85, 2018.

[9] B. J. Frey and D. Dueck. Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972–976, 2007.

[10] W. Ge, X. Lin, and Y. Yu. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3034–3043, 2019.

[11] B. G. Gebre, P. Wittenburg, and T. Heskes. The gesturer is the speaker. In *ICASSP*, pages 3751–3755, 2013.

[12] Binyam Gebrekidan Gebre, Peter Wittenburg, Tom Heskes, and Sebastian Drude. Motion history images for online speaker/signer diarization. In *ICASSP*, pages 1537–1541, 2014.

[13] J. M. Górriz, J. Ramírez, E. W. Lang, C. G. Puntonet, and I. Turias. Improved likelihood ratio test based voice activity detector applied to speech recognition. *Speech Commun.*, 52(7–8):664–677, 2010.

[14] F. Haider, N. Campbell, and S. Luz. Active speaker detection in human machine multiparty dialogue using visual prosody information. In *IEEE Global Conference on Signal and Information Processing*, pages 1207–1211, 2016.

[15] K. Hoover, S. Chaudhuri, C. Pantofaru, M. Slaney, and I. Sturdy. Putting a face to the voice: Fusing audio and visual signals across a video to determine speakers. *CoRR*, abs/1706.00079, 2017.

[16] H. Hung and S. O. Ba. Speech/non-speech detection in meetings from automatically extracted low resolution visual features. In *ICASSP*, 2010.

[17] B. Joosten, E. Postma, and E. Krahmer. Voice activity detection based on facial movement. *Journal on Multimodal User Interfaces*, 9(3):183–193, 2015.

[18] D. Klotz, J. Wienke, J. Peltason, B. Wrede, S. Wrede, V. Khalidov, and J.-M. Odobez. Engagement-based multi-party dialog with a humanoid robot. In *SIGDIAL Conference*, page 341–343, USA, 2011. Association for Computational Linguistics.

[19] Q. Liu, W. Wang, and P. Jackson. A visual voice activity detection method with adaboosting. In *Sensor Signal Processing for Defence*, pages 1–5, 2011.

[20] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[21] G. Monaci. Towards real-time audiovisual speaker localization. In *2011 19th European Signal Processing Conference*, pages 1055–1059, 2011.

[22] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[23] P. Liu and Z. Wang. Voice activity detection using visual information. In *ICASSP*, volume 1, pages I–609, 2004.

[24] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, page 1742–1750, 2015.

[25] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi, and C. Pantofaru. Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4492–4496, 2020.

[26] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE ICCV*, pages 618–626, 2017.

[27] M. Shahid, C. Beyan, and V. Murino. Comparisons of visual activity primitives for voice activity detection. In *Image Analysis and Processing (ICIAP), Lecture Notes in Computer Science, vol. 11751, E. Ricci, S. Rota Bulo, C. Snoek, O. Lanz, S. Messelodi, N. Sebe (eds), Springer, Cham,*, pages 48–59, 2019.

[28] M. Shahid, C. Beyan, and V. Murino. Voice activity detection by upper body motion analysis and unsupervised domain adaptation. In *ICCVW*, 2019.

[29] R. Sharma, K. Somandepalli, and S. Narayanan. Toward visual voice activity detection for unconstrained videos. In *IEEE ICIP*, pages 2991–2995, 2019.

[30] S. Shum, N. Dehak, E. Chuangsuwanich, D. A. Reynolds, and J. R. Glass. Exploiting intra-conversation variability for speaker diarization. In *INTERSPEECH*, 2011.

[31] D. Sodoyer, B. Rivet, L. Girin, C. Savariaux, J.-L. Schwartz, and C. Jutten. A study of lip movements during spontaneous dialog and its application to voice activity detection. *Journal of the Acoustical Society of America*, 125(2):1184–1196, 2009.

[32] K. Stefanov, J. Beskow, and Giampiero Salvi. Vision-based active speaker detection in multiparty interaction. In *Int. Workshop Grounding Language Understanding*, pages 47–51, 2017.

[33] F. Tao and C. Busso. End-to-end audiovisual speech activity detection with bimodal recurrent neural models. *CoRR*, abs/1809.04553, 2018.

[34] E Verteletskaya and Kirill Sakhnov. Voice activity detection for speech enhancement applications. *Acta Polytechnica*, 50(4), 2010.

[35] W. Yunchao, L. Xiaodan, C. Yunpeng, S. Xiaohui, C. Ming-Ming, F. Jiashi, Z. Yao, and Y. Shuicheng. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2314–2320, 2017.

[36] X. Zhang, Y. Wei, J. Feng, Yi Yang, and T. S. Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2018.

[37] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, and B. Schuller. Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition. *IEEE Access*, 7:97515–97525, 2019.