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

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

***Abstract*—We present an automatic voice activity detection (VAD) method that is solely based on visual cues. Unlike traditional approaches processing audio, we show that upper body motion analysis is desirable for the VAD task. The proposed method consists of components for body motion representation, feature extraction from a Convolutional Neural Network (CNN) architecture and unsupervised domain adaptation. The body motion representations as images are used by the feature extraction component, which is generic and person-invariant, thus, can be applied to a subject who has never been seen. The endmost component handles the domain-shift problem, which appears due to the fact that the way people move/ gesticulate while speaking might vary from subject to subject, which results in disparate body motion features and consequently poorer VAD performance. The experimental analyses applied on a publicly available real-world VAD dataset show that the proposed method performs better than the state-of-the-art video-only and multimodal VAD approaches. Moreover, the proposed method has a better generalization ability as VAD results are more consistent across different subjects. As another major contribution, we present a new multimodal dataset (called RealVAD), created from a real-world (no role-plays) panel discussion. This dataset contains many actual situations/ challenges that are missing in the previous VAD datasets. We benchmarked the RealVAD dataset by applying the proposed method as well as cross-dataset analyses. Particularly, the results of cross-dataset experiments highlight the remarkable positive contribution of the unsupervised domain adaptation applied.**

***Index Terms*—voice activity detection, active speaker, body motion analysis, nonverbal behavior, visual cues, real-world dataset, unsupervised domain adaptation.**

I. Introduction

The task of automatically detecting "Who is Speaking and When" by processing audio data only, video data only or these two modalities together, is broadly named as Voice Activity Detection (VAD) [1]. Automatic VAD is a very important task and also the foundation of several domains, e.g., human-human, human-computer/ robot/ virtual-agent interaction analyses, and industrial applications. For instance, VAD has been applied for content analysis of the conversations during human-computer interactions (HCI) in [2]. The robust detection of active speaker is the key to reply to a specific interlocutor during human-robot interactions in [3]. Furthermore, the results of VAD have been used to extract nonverbal features to analyze the human-human social interactions in [4], [5]. Differently, in [6], a VAD approach was applied to isolate the respiration segments for the classification of breathing movements. In panel discussions, public presentations, debates, etc., the camera is manually focused on the person who is speaking. This in fact can be automated using an effective VAD technology. In entertainment industry, VAD systems have been used to place the transcriptions on a location close to the active speaker, which makes reading the subtitles more feasible as one can still focus on the scene as well [7]. The contributions of VAD is not restricted with these applications and can be extended to video retrieval, person specific model adaptation for speaker recognition [8], and so forth.

Typically, VAD has been performed by processing audio signal only and there is a wide range of literature regarding that. However, carrying out audio-based VAD can be challenging, e.g., in unstructured social gatherings, cocktail party scenarios, and especially when the number of subjects is high or when speakers are located close to each other. Recently, multimodal VAD [9], [10] is gaining a lot of attraction. Multimodal VAD approaches are either based on the synchronization of audio and video or by joint modeling of these two modalities [10] such that the visual information is used to associate the active subject to a speech [2]. As compared to single modality-based VAD, multimodal VAD might result in more accurate performance. However, it might not be applicable when the audio data is not available, e.g., due to technical, ethical or legal issues. In such cases, video only VAD (referred to as "visual VAD" for the rest of this paper) is very desirable. There are relatively less study examining visual VAD (VVAD) and thus, better performing methodologies are needed.

A major contribution of this paper is to introduce a novel VVAD methodology. To date, the majority of the VVAD approaches have attempted to detect facial motion, e.g., [11] or specifically lips motion, e.g., [12]–[15]. There are also few works considering other visual cues, e.g., head activity [8], hand movements [8], [15], [16], and visual focus of attention (VFOA) [16]. In this study, we analyze the upper body motion of a person "holistically", i.e., without relying on the detection of a specific body part. Using body motion-based cues to detect voice activity is motivated from several findings. Some examples are quoted as follows. In [17], it is shown that during social interactions individuals unintentionally synchronize their nonverbal and linguistic behavior. In [18], it is proved that gestures occur mainly during speech (with a delay of milliseconds). In [19], analyses showed that, during narrations,

* 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).

about 90% of the time all gestures occur while a person is speaking. It was observed in [20] that, during a conversation, the occurrence of body activity while speaking is for more than 80% of the total speaking time. Additionally, several approaches have exploited the relationship between speech and body motion cues for various applications. For instance, in [21], the synchronization between pitch and gestures were used to obtain more human-like artificial agents. In [22], the body motion cues were used to estimate emergent leadership, and to detect the personality traits.

The proposed method extracts upper body motion cues using a Convolutional Neural Network (CNN), trained once, and can be applied to any (new) person without requiring re-training. Therefore, it supports an end-to-end training, where motion cues are learned from data itself. According to psychology literature, the way people move/ gesticulate while speaking might vary a lot from subject to subject. This possibility introduces the challenge of domain specificity (so-called domain-shift problem) for automatic VAD. In detail, the distribution of training data and the test data belonging to different subjects can be atypical. This results in dissimilar body motion representations and consequently poor VAD performance. Surprisingly, domain-shift problem has not been addressed by any previous VAD study before. Herein, we tackle this problem by adapting a simple but effective domain adaptation technique, which does not require labels belonging to the test data (unsupervised), and supports person-invariant training. In other words, the proposed domain adaptation technique does not learn person-specific features, but provides a common effective representation between the training domain and the new person's domain, resulting in larger generalization ability.

Another major contribution of this study is to present a new real-world multimodal VAD dataset, named RealVAD. Independent to the modality being used, many VAD approaches in the literature achieved desired performances. However, most of the time, they have been tested on datasets composed of constrained scenarios or role-plays, which are limited to include many real-world situations. Even very common circumstances, for instance; occlusions and background motion were not involved/ avoided during the creation of some of the VAD datasets. Besides, the majority of the datasets have a VAD ground-truth constituted by an automatic speaker diarization method (the process of labeling a speech signal with labels corresponding to the identity of speakers) that might result in noisy/ incorrect annotations. Moreover, there is no publicly available real-world VAD dataset having more than three subjects at a time. Another important issue that existing VAD datasets have not addressed, is containing persons from different ethnic origins. Having high number of persons at a time and ethnic diversity are important criteria as they potentially increase the human behavior diversity and consequently, allow researchers to evaluate their method in a wider perspective.

RealVAD dataset addresses all (and even more) these aforementioned real-world situations. It was constituted from a YouTube video that can be found in [23], which contains a panel discussion. All the properties of RealVAD including the annotation process are given in Section III and supplementary material. The dataset[1] includes the upper body detections of nine panelists (bounding boxes; the location of panelists' upper body in a video frame), which were manually extracted, the corresponding video frame number, speaking status (speaking or not-speaking) determined as a result of two human annotators' audiovisual inspection, and acoustic features (Mel-frequency cepstral coefficients (MFCC) and raw filterbank energies) that were automatically extracted. RealVAD dataset is particularly suitable to test VVAD approaches based on the upper body motion as this information is already extracted, made publicly available and the corresponding features were benchmarked by applying the proposed method. In the meantime, it allows researchers to investigate other cues based on any other modality for VAD task.

The contributions of this paper can be summarized as follows.

- We propose a novel visual VAD (VVAD) method, which performs not only better than the state-of-the-art (SOA) VVAD methods but also better than the SOA multimodal VAD approach. This is also the first attempt that unsupervised domain adaptation is integrated to VAD. It is shown that domain adaptation provides not only better performance but also more consistent VAD results such that the detection performances are equally well for all subjects.
- We introduce a new multimodal real-world dataset (RealVAD), which includes many circumstance, e.g.; having persons from different ethnic origins, including higher number of subjects at a time, containing person detections with different (sometimes) non-static backgrounds, having natural changes of illumination and shadows that make motion analysis harder, having highly imbalanced VAD ground-truth, and so forth.
- We benchmark RealVAD dataset by applying the proposed method, and also by performing other analyses including cross-dataset experiments that allow us to capture more insights about the characteristics of the datasets used. Especially, the results of cross-dataset analyses prove the remarkable positive contribution of applying unsupervised domain adaptation.
- We conduct an extensive literature review regarding V-VAD approaches and the VAD methods performing body motion analysis.
- We present a comparative study of the existing VAD datasets, and also define some design criteria that are useful for the creation of future VAD datasets.

The rest of this paper is organized as follows. A discussion including the previous computer vision-based VAD approaches, multimodal VAD methods using body motion analysis, and an evaluation of the existing VAD datasets, are given in Section II. RealVAD dataset is explained in Section III. In Section IV, the details of the proposed methodology are described. The experimental analyses with results not only for RealVAD dataset but also for an existing real-world dataset, are given in Section V. This section also includes cross-dataset experiments and the real-time application of the proposed

[1]link will be inserted after acceptance of this paper

method. Finally, in Section VI, we conclude the paper with a summary and discussions including the possible future investigations that can be performed regarding the proposed method and RealVAD.

II. Related Work

In this section, visual VAD (VVAD) methods and other VAD approaches (e.g., multimodal) using body motion analysis, are discussed. The existing VAD datasets are also reviewed and evaluated in terms of some design criteria defined.

A. Visual VAD

Video only VAD (visual VAD, VVAD) methods can be categorized in terms of the body parts examined: *i)* facial cues-based (e.g., facial landmarks movement, lips motion), *ii)* body cues-based (e.g., hand gestures, head movement, upper body motion) or *iii)* a composition of these two.

As a preliminary study, in [12], a Bayes Net model processing the results of face detection, skin color and texture detection and mouth motion sensors to determine the speaking status of a person in a human-machine interaction platform was presented. Similarly, Joosten et al. [13] used Spatiotemporal Gabor filters to extract head, mouth and lips motion and came up with a conclusion that the mouth motion performs the best out of all for VAD task. Following that, temporal orofacial features have been explored in [24] such that Principal Component Analysis (PCA) was applied to merge different features that are input to expectation-maximization algorithm defining speech and not-speech regions. In [14] head and lips displacement cues have been used to localize the active speaker for a human-machine multiparty interaction application. Detailed analyses of head movements versus the fusion of head and lips movements were performed while the model was trained in speaker dependent or independent ways. In that study [14], it was shown that the contribution of the head motion is significant for speaker independent setting, while the combination of head and lips movements is the best performing. More recently, Stefanov et al. [11] have utilized face features extracted from AlexNet [25] to perform VAD in a real-time multiparty interaction setting. Long short-term memory (LSTM) has been used to model the temporal information between face features over time, which also predicts whether a given frame composed of only face is speaking or not. As an extension of that work [11], authors addressed VVAD in a multi-person language learning scenario, which the auditory modality is fundamental [26]. The main difference between [11] and [26] are: using VGG-16 [27] instead of AlexNet [25] and investigating the effect of using temporal models by comparing the performance of LSTM models to non-temporal Perceptron models.

Body cues, specifically, hand gestures and upper body motion analysis have been applied for VVAD as well. Hung et al. [16] proposed a method using hand activity and VFOA to perform VAD in small group meetings. The motivation behind using these cues was that the speaker is the one who moves the most and the majority of the subjects should be looking towards the speaker. In that study [16], different supervised and unsupervised machine learning techniques have been compared, showing promising results particularly for the utilization of VFOA. Later, by using the same dataset, Gebre et al. [28] tested motion history images, which showed significant performance as well. By using CCTV data, Cristani et al. [1] have evaluated the correlation between gestures and speech aiming to detect the speaker. The feature descriptor used in that study [1] is based on the optical flow of body encoded as the optical flow energy and complexity. The dataset used in that study [1] can be seen more unconstrained as compared to the meeting dataset in [16], [28]. However, [1] includes only the top-view, which already reduces the chances of occlusions, and the frames, which the area of interests of people are overlapping, were discarded from the evaluation of their method. Chakravarty et al. [15] have used directional audio information to annotate the speaking status of each person while the motion in head and shoulders is represented in terms of the improved trajectory features (ITF) [29]. In their further work, the authors used ITF to perform person-specific VVAD with an online learning mechanism [30].

B. Other VAD Approaches Using Body Motion Analysis

A pioneer audiovisual VAD method, which considered analyzing body motion is [31]. It has been tested on human-centered user-interfaces such that face, skin, texture, lips motion, and silence detectors have been optimally fused with the contextual information using a Dynamic Bayesian Network architecture that exploits the temporal correlation between audio and visual sensors.

Another work using body motion together with speech features was [9], evaluated on a dataset composed of meetings recorded using a stationary camera and a single microphone. That approach has exploited the long-term co-occurrences between audio and video subspaces found by clustering. It does not rely on a priori about speakers for training. However, it is not clear whether that method [9] is able to detect simultaneous active speakers. Friedland et al. [32] combined MFCC-based acoustic features with body motion cues, which are represented as the average motion vector magnitude in skin blocks. That multimodal method showed enhanced performance as compared to the audio-only baseline described in the same study [32]. Later on, in [33], person-specific models have been learned using speech samples corresponding to the gestures. The occurrence of gestures indicated the presence of speech and the location of gestures indicated the identity of the speaker.

Graphical models have been commonly adapted for VAD as well, such that the speaking status of a person is determined using audio information only while body motion features is used to estimate the position of the active speaker. For instance, in [2], multichannel audio signal has been used to get direction information along with a visual tracker to affiliate multiple audio signals to multiple persons. On the other hand, Darca et al. [34] presented a hierarchical audiovisual system applied to surveillance scenarios that detects the active speaker by tracking them and recognizing their voices. Multiple challenges including large occlusion and cross-talks

were overcome with the help of using multiple modalities. However, that approach [34] is limited as only the most dominant speaker is detected and tracked while the other persons speaking simultaneously with the dominant speaker, cannot be detected.

As a different approach, in [35], cross-modal supervision-based VAD has been addressed. In detail, video is used within an audiovisual co-training such that a generic body cues-based VAD classifier trained by directional audio, which is used to train a video-based person-specific VAD. The trained video classifier is used to supervise the training of personalized voice models. The only shortcoming of that method [35] is being person-specific, thus, requiring training data for each new person. Recently, Hoover et al. [36] has used speech represented as spectrograms and visual features extracted from facial features. The speech and face embeddings were jointly modeled with a CNN and a bidirectional LSTM model to perform speaker-independent VAD. The advantage of that method [36] is not requiring any prior information about the number of persons. However, it [36] requires accurate pre-trained face and speech detectors. Tao et al. [37] have proposed a bimodal recurrent neural network (BRNN) framework, which consists of three RNNS: audio (A-RNN), visual (V-RNN) and audiovisual (AV-RNN). The A-RNN includes fully connected (FC) and LSTM layers to process acoustic Mel-filterbank features. The V-RNN includes a CNN extracting a visual representation and LSTMs to process temporal information. The AV-RNN relies on fully connected layer and LSTM to process the concatenated output from A-RNN and V-RNN. BRNN is jointly trained by minimizing a common cross-entropy function. Chung et al. [10] investigated a two-stream CNN model to learn the sound and lips motion embeddings within a single model. The results shown in [10] proves that joint learning of audio and synchronized lips motion could improve the active speaker detection results as compared to the body motion-based VAD presented in [30].

The VAD task becomes very challenging when there are many persons in an unstructured environment such as in a mingling scenario. Different from any work discussed in this paper, Gedik et al. [38] used the data coming from triaxial accelerometer sensors worn around the neck of subjects to perform VAD, which are evaluated in a very crowded scenario. In that work [38], the power spectral density of the motion signal was utilized as the feature representation of the body evolution while the model training was person-specific. For the first time, in [39], the depth visual information has been fused with audio and planar video information to perform VAD. The results showed that depth information significantly contributes to VAD. However, that study [39] is limited as being only tested in simple scenarios having two persons.

A summary of the studies discussed in Section II-A and II-B are given in Table I. That summary is in terms of the scenarios these works have been evaluated on and the audio/ video/ audiovisual cues they have utilized.

TABLE I: The summary of VAD approaches in terms of the audio/ visual cues they have utilized, and the scenario types they have been evaluated on.

Scenario	Audio/Visual Cues
Video surveillance	Body motion [1]; MFCC [34].
Small group meetings/ Dyadic interactions	MFCC [9], [10], [32], [37]; spectograms [36]; temporal orofacial features [24]; body motion [9], [28], [32], [33], [39]; face motion [13], [36]; lips motion [10], [13], [37], [39]; hand activity and VFOA [16]; head motion [13]; depth information [39].
Human-machine interaction	Face motion [12], [31]; deep face features [11], [26]; lips motion [12], [14], [31]; head motion [2], [14]; torso motion [2]; short-time Fourier transform of audio signal [2].
Mingling scenarios	Power spectral density of the motion signal [38].
Panel/ Presentation	Upper body (head and shoulders only) motion [15], [30], [35].

C. The Existing VAD Datasets

There are many VAD datasets, which were specifically collected to perform this task. In addition to them, some corpora that were built for various other purposes have been also used to evaluate VAD methodologies. This section reviews the ones presented in the last decade.

AMI [40] is a multimodal, publicly available dataset, composed of role-plays in meeting environments. It has been used for various purposes while VAD is one of them. Each video of it includes one subject's upper body, captured by a close-up camera from frontal view when the background is static. There are no occlusions among subjects. Canal9 [41] is a multimodal database of political debates for the analysis of social interactions. It is composed of role-plays, thus, it is not a real-world dataset. Majority of the time, it has been used for the automatic detection of conflicts in spoken conversations or for the automatic detection of (dis)agreements between persons from nonverbal behaviors. It was also utilized to test various VAD approaches, which are mostly audio based only. It includes a large number of subjects (in total 190). In some of the shots, subjects are not close to each other, thus, there is no occlusions among them. For all type of shots, only the upper body of the subjects is visible. It contains a single static background in all shots. AVDIAR [42] is publicly available another multimodal role-play dataset, composed of 23 video sequences ranging from ten seconds to three minutes. Its VAD annotation is for 27 minutes. Similar to AMI [40] and Canal9 [41] datasets, it contains multiparty dialogues among speakers. However, it is composed of relatively simpler scenarios, which are limited to reflect many real-world situations when subjects are either static or moving. Its visual data annotation is composed of bounding boxes of faces and upper-body positions at each video frame. It includes video sequences having more than one subject at a time (up to three subjects) but there are few frames where partial occlusions among subjects occur.

As an example of audiovisual person tracking datasets that are used for the evaluation of VAD methods (not VVAD but audio-based and/or audiovisual VAD) as well, RAVEL [43] can be given. It is composed of role-plays of human-robot interaction scenarios, e.g., a person talking on the phone, commanding a robot, asking the robot for instructions, human introducing a new person (or vice versa), and cocktail party. Half of its scenarios include only one person where

there is no simultaneous speech. The cocktail party scenario includes the maximum number of subjects at a time (five persons) where partial occlusions happen. The scenarios are very short, relatively simple and not always performed very naturally. All recordings have the same static background. That publicly available dataset's annotation [43] includes the 2D/3D positions of the subjects. However, neither separated audio from audiovisual recordings nor extracted audio features are available. Similarly, AVTRACK-1 dataset [44] is composed of role-play multiparty conversations, proposed for audiovisual person tracking and also used for VAD. Only image-plane annotations of it are publicly available. Some recordings of it include partial to full occlusions. It is a very small scale dataset (composed of in total four videos, each lasts approx. 20 seconds), with few subjects (max. three people at a time) who move slowly in the camera field of the view, are close to the camera, and mostly facing it when the background is static. AVASM [45] is another publicly available multimodal dataset, composed of role-plays for audiovisual person tracking. It includes simple scenarios such as subjects counting numbers. It is more suitable to investigate audio-based VAD as its recordings have relatively narrow camera field-of-the-view, where subjects move (without replacement) in a very limited space, and usually look at the camera. Additionally, it targets more to be a benchmark for testing speaker diarization methods by including continuous simultaneous speakers. However, the sound source is a stationary loudspeaker that emits white noise and speech from different positions, while VAD ground-truth is based on an automatic speaker diarization method that might result in noisy/ erroneous labels. That dataset [45] includes maximum two persons at a time while videos are short and all were captured with the same static background. Only image-plane annotations of AVASM dataset [45] are available.

On the other hand, the focus of MVAD dataset [46] is HCI systems. It was captured with a setting of a microphone array, a color camera, and a depth sensor. It is composed of role-plays with very simple scenarios having short silent segments. All subjects are from the same ethnic origin. VAD ground-truth is based on an automatic speaker diarization technique, thus labels can be noisy/ erroneous. The subjects always gaze the camera and are not moving much (no replacements) through all the recordings while background is not changing. That dataset is composed of short videos (40-60 seconds) and only image-plane annotations corresponding to VAD ground-truth are publicly available. Recently, CAV3D [47], which is a public multimodal dataset is presented. It contains in total 20 role-play short video sequences such that nine of them are with a single speaker; six of them are with a single active speaker and a second person who is not speaking; and five of them are with simultaneous speakers up to in total three persons. The subjects in some of the recordings are moving, thus, occlusions might occur. However, they are always looking at the camera, which makes the recordings very unrealistic. Its visual data annotations are limited to mouth positions that last 15 to 80 seconds. VAD ground-truth of it is obtained by using a tool for segmenting, labeling and transcribing speech, thus might be noisy/ erroneous.

Another very recent dataset; AVA-ActiveSpeaker [48] provides VAD ground-truth for 15-minutes long movie clips, which were labeled by three annotators. The VAD annotations were validated by applying Fleiss kappa test, resulting in high reliability. That dataset contains speaking activity (speaking and audible, speaking but not audible, and not speaking) associated with a visible face extracted by using automatic face detection and tracking algorithms. In total 3.65 million faces were obtained from continuous segments of 160 movies on YouTube. Even though, most of the time whole upper body and/or whole body of the speaker is visible only face is annotated with a VAD label. Movies are with high resolution and from film industries around the world, leading to diversity in recording conditions, and speaker demographics. Separated audio and/or audio-based features are not provided. Some clips include simultaneous talks and turn takings while some has monologues. The dataset includes different environments, thus, background is various, camera motion exists, context and behaviors are diverse. There are some crowded scenes as well. However, this does not mean that such crowded scenes include VAD/ face annotations of each person in the field of the view of the camera. In detail, annotations were mainly performed for the speakers specially if and only if the face of her was successfully detected and tracked. Given that in movies camera more focuses on the speakers to better reflect the nonverbal behaviors and emotions of the actors, the annotated faces are usually with high resolution (very close to camera) and not occluded. As compared to real-world situations, movie clips are still very controlled, but as they are performed by professionals, the behaviors of persons are more realistic than other aforementioned role-play VAD datasets.

Unlike any previously mentioned dataset in this section, MS.Thesis [15] is from a real-world recording. It contains the examiners' movements while asking questions during thesis defense presentations. It is composed of seven videos having the same three examiners when only the upper bodies of them are visible. Its VAD ground-truth was obtained with an automatic speaker diarization method, thus, might be noisy/ erroneous. Unfortunately, it is currently not publicly available. On the other hand, the same authors [15] presented Columbia dataset [30], which is composed of a single panel discussion video. Excluding ours, it is the only publicly available real-world VAD dataset. It has been used to test visual and audio-visual VAD approaches. Its VAD ground-truth is 35 minutes only and was obtained by applying an automatic speaker diarization method (can be noisy/ erroneous). It includes three panelists at a time while in total five panelists exist. Therefore, each panelist has VAD annotations less than 35 minutes. Panelists never talk simultaneously. Only the visual data, i.e., the locations of subjects' head including shoulders are supplied while separated audio and/or audio-based features are not provided. It includes the same background for all panelists while in a couple of frames camera motion exists.

As seen, the number of real-world datasets is very limited. The datasets are usually elicited in a lab setting where the subjects may not behave naturally as they perform the imposed role-play instead of spontaneous behaviors. Additionally, majority of the datasets are called multimodal but there are cases

the separated audio signal and/or audio based features are not provided, e.g., [15], [30], [43], [45], [46], [48] or visual data annotations are available only for some body parts, e.g., face only [30], [48], lips only [47]. For the creation of a big dataset, it is convenient to extract audio/video features automatically (e.g., in [48] the location of faces were annotated by applying automatic face detection and tracking algorithms), but this might result in false positives. Especially, VAD ground-truths that were obtained by applying basic speaker diarization methods, e.g., as applied in [15], [30], [45], [46] might contain noisy/ erroneous labels.

The datasets discussed in this section and our proposed dataset (RealVAD) are evaluated in terms of the following criteria in Table II.
C1: Not being based on role-plays.
C2: Having multiple subjects at a time in the field of the view of the camera.
C3: Having VAD annotations of simultaneous speakers.
C4: Having subjects with different ethnic origins and/or diversity in demographics.
C5: Having non-static background and/or camera motion.
C6: Having diversity in video resolutions and/or in the distance between camera and the subjects.
C7: Having natural changes of illumination and/or shadow.
C8: Having partial/full occlusions between subjects.
C9: Providing publicly available VAD ground-truth.
C10: Providing VAD ground-truth labeled by more than one human annotators and showing the reliability of annotations, i.e., high agreement between annotators.
C11: Providing publicly available annotated visual data (e.g., location of the subjects/faces in a video frame).
C12: Providing publicly available separated audio data (e.g., output of a microphone array, etc.) and/or extracted acoustic features (e.g., MFCC).
From C1-C8, the more criteria a dataset satisfies means that it more includes the complex dynamics of the real-world. In addition to that, satisfying C9-C12 means that the corresponding dataset is a good candidate to be used for the testing of VAD approaches.

TABLE II: The evaluation of existing VAD datasets and the proposed dataset (ReadlVAD) using criteria C1-C12 (see text for the description). NM stands for not mentioned.

DATASET	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
AMI [40]	✗	✗(=1)	✓	NM	✗	✗	✓	✗	✗	✗	✗	✓
Canal9 [41]	✗	✓(≤5)	✓	NM	✗	✓	✓	✓	✓	✓	✗	✓
RAVEL [43]	✗	✓(≤5)	✓	NM	✗	✓	✓	✓	✓	✗	✓	✗
AVTRACK-1 [44]	✗	✓(≤3)	✓	NM	✗	✓	✓	✓	✓	✗	✓	✓
AVASM [45]	✗	✓(≤2)	✓	NM	✓	✗	✓	✗	✓	✗	✓	✓
MVAD [46]	✗	✓(=3)	✓	✗	✗	✗	✓	✗	✓	✗	✓	✗
MS.Thesis [15]	✓	✓(=3)	✗	NM	✗	✗	✓	✗	✗	✗	✗	✗
Columbia [30]	✓	✓(=3)	✗	NM	✓	✗	✓	✗	✓	✗	✓	✗
AVDIAR [42]	✗	✓(≤3)	✓	NM	✗	✓	✓	✗	✓	✗	✓	✓
CAV3D [47]	✗	✓(≤3)	✓	NM	✗	✓	✓	✓	✓	✗	✓	✗
AVA [48]	✗	✓(≤6)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
RealVAD	✓	✓(=9)	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓

III. REALVAD DATASET

RealVAD dataset is constructed from a YouTube video composed of a panel discussion lasting approx. 83 minutes [23]. The audio is available from a single channel. There is one static camera capturing all panelists, the moderator and audiences. The field-of-the-view of the camera while the panelists are shown with the assigned identification numbers are given in Figure 1. A particular advantage of RealVAD

Fig. 1: A sample frame from the video that RealVAD dataset is constituted from. The identifier numbers (one to nine) are associated to the panelists below them.

is to allow studying the effect of ethnic origin variety to the automatic VAD as it is composed of panelists with different nationalities (British, Dutch, French, German, Italian, American, Mexican, Columbian, Thai). In this study, we have not investigated this topic much. However, the unsupervised domain adaptation method applied showed better results as compared to not applying it. There is a gender balance such that there are four female and five male panelists.

Panelists are sitting in two-rows. They are not looking at the camera unlike it frequently happens in other datasets. They can be gazing audience, other panelists, their laptop, the moderator or anywhere in the room while speaking or not-speaking. Therefore, they were captured not only from frontal-view but also from side-view varying based on their instant posture and head orientation. They are moving freely and are 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). Given that the position of the camera is static, the distance between it and each panelist is also varied. The background of the front and back row panelists are also different. Especially, 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.

A. Annotation of RealVAD Dataset

We provide acoustic features (MFCC and raw filterbank energies) that were automatically extracted, visual features (the upper body detections) that were manually defined and the

corresponding VAD ground-truth for nine panelists. The VAD ground-truth was determined by human annotators by performing a two-stage inspection. Former stage was performed using the audiovisual recording and the further stage was based on visual data only.

In detail, first, by using ELAN [49] annotation tool, two annotators separately labeled the speaking activity (speaking or not-speaking) in a time-stamp. This was performed considering only the panelists, i.e., the moderator was not considered, and resulted in the annotations representing whether there is a speaker or not at a time but not the speaker's identity. In other words, in this stage of the VAD annotation, we did not use person detections, such that whole audiovisual recording was used. This was convenient as there are no simultaneous speakers. This first stage allowed us to detect the frames that only moderator is speaking, which decreased the labeling time in the second stage as the frames that moderator is speaking were already retrieved as not-speaking for all panelists, and also helped annotators to be familiar with the labeling task.

On the other hand, for each frame, the person detections were extracted by using human annotators, i.e., without relying on an automatic detection algorithm. Person detections are the bounding boxes, each covering the upper body of a single panelist with background. The resolution of them can be diverse even for the same person (for different postures of a panelist, her bounding box can be bigger or smaller given that we want to include all upper body parts), which might introduce an additional challenge. The corresponding MFCC and raw filterbank energies that were extracted automatically are also publicly available.

As the second stage of VAD annotation, the same two annotators once again separately but this time by using only the person detections (bounding boxes belonging to only one panelist at a time), designated the VAD labels. During this labeling, annotators were able to see their previous annotations. In detail, as an example; during the labeling process of panelist number one (1), 30 consecutive bounding boxes (equals to one second) were played at each time without overlapping. If these bounding boxes were from the frames moderator is speaking according to the first stage of the annotation, then the suggested label was shown as zero to the annotator, otherwise it was shown as minus one (-1). Once the annotator entered the VAD label, if it was one (speaking), then for the other eight panelists, we saved the VAD label as zero, which were retrieved as the suggested labels during their annotation. The user interface used for the second stage of VAD annotation is given in supplementary material. Once two annotators finalized all labeling task, the same bounding boxes having different labels were discarded. Therefore, the VAD ground-truth of this dataset has a perfect (100%) agreement.

We have noticed that the dataset composed after VAD annotation is highly imbalanced such that the number of not-speaking bounding boxes is much more than the number of speaking bounding boxes. This is as a result of having eight not-speaking panelists while the remaining one panelist is speaking, which occurs nine times, i.e., for each panelist's turn. However, having so many not-speaking bounding boxes does not contribute to the behavior diversity much, and conversely, introducing redundancy. Therefore, we removed the annotation of some not-speaking bounding boxes from the ground-truth by considering the ratio between the total number of speaking and not-speaking bounding boxes per panelists (see supplementary material for more information). This removal first performed on the very short segments, which were appeared due to discarding the bounding boxes having annotator disagreement. Our dataset still includes 4.5 times more not-speaking bounding boxes than speaking bounding boxes. We have preferred to keep that ratio since it introduces a new challenge for training a classifier given that having an imbalanced data can be misleading (see [50] for more details).

Overall, the total number of annotated video frames is 74625, which corresponds to 671625 bounding boxes. The agreement between annotators in terms of Cohens-Kappa coefficient is 0.71 (substantial agreement [51]) before discarding any bounding boxes. After discarding some not-speaking bounding boxes as mentioned above and the bounding boxes labeled differently by annotators, the final number of bounding boxes with VAD ground-truth is 380658, which corresponds to 74327 video frames (approx. 42 minutes). For each panelist, the frame number, the x and y coordinates of left-top corner of the bounding box, the width and height of the bounding box, the corresponding VAD label and the acoustic features are publicly available. More information (e.g., the number of discarded bounding boxes per panelist, bounding boxes showing freely moving panelists, i.e., doing various actions resulting in different postures, bounding boxes with background motion and occlusions) regarding RealVAD dataset can be found in supplementary material.

IV. Proposed Methodology

The proposed method is illustrated in Figure 2. First, multiple dynamic images [52] are generated from successive RGB frames and they are divided into sub-dynamic images, each covers a single person's upper body and the background (Section IV-A). This is performed for the training and test streams individually. Then, ResNet50 [53] is fine-tuned for VAD task when the sub-dynamic images of the training stream are the inputs. This fine-tuning (Section IV-B) is applied in two ways: *i)* a fully connected layer called as $Fc1$ in Figure 2 is added to the existing ResNet50 architecture and the weights of Fc and $Fc1$ are updated or *ii)* without adding any new layer, the weights of the convolutional layers of ResNet50 (Block 4 and 5 in Figure 2) as well as Fc are updated.

Once ResNet50 is fine-tuned, features are extracted from $Fc1$ layer in case fine-tuning option *(i)* is applied or Fc layer when fine-tuning option *(ii)* is applied, for each training and test sub-dynamic image. These features are given to the unsupervised domain adaptation component (Section IV-C). Unsupervised domain adaptation and Support Vector Machine (SVM) learning (Section IV-D) are jointly performed. The domain adaptation method aligns the subspace of training (source domain) and test (target domain) features after applying two-layer stacked sparse autoencoder. This results in new feature representations for training and test streams. Meanwhile, an SVM classifier is learned with the new feature representations of the training stream and the corresponding VAD labels.

To infer the VAD label (speaking or not-speaking) of the test stream, its new feature representations (i.e., domain adapted features) are given to the trained SVM classifier. Each predicted VAD label corresponds to a specific person in the test streams, those the test sub-dynamic image is constructed from.

A. Construction of Multiple Dynamic Images

There are various ways to detect and represent the body motion of a person. One of the most popular methods is optical flow. On the other hand, Bilen et al. [52] presented a very effective method, which is called dynamic image representation. That methodology [52] achieved the SOA activity recognition results. Dynamic image (DI) representation summarizes the short-term spatio-temporal content of a video in a single image. It can be seen as a compact representation of a video segment, which summarizes the appearance and motion of it. Construction of a DI contains rank-pooling that encodes the temporal evolution of the frames in a video and potentially enables the use of any CNN model with fine-tuning. The details of its algorithm can be found in [52].

In our previous work [54], we have compared different optical flow image representations, DI constructed from RGB images and DI constructed from optical flow images for VVAD. The results showed that RGB based DIs are one of the most effective. Motivated by the results given in [54], we obtain one DI from 10 consecutive RGB frames (no frame overlapping).

A DI is, then divided into sub-dynamic images such that each of them contains a single person's upper body and the background. In detail, for the experiments performed on the RealVAD dataset, the region of a sub-dynamic image was determined according to the bounding box annotations supplied by the dataset. For the experiments applied on the Columbia dataset, we extracted the bounding boxes such that they contain the whole upper body of a panelist and obtained the sub-dynamic images with repect to these bounding boxes.

B. ResNet50 Training and Feature Extraction

Given that dynamic images can be used with any CNN architectures for fine-tuning [52], we first used AlexNet [25] (pre-trained on ImageNet dataset). However, this resulted in inferior performance due to overfitting (refers to experiments with Columbia dataset [30]) even regularization techniques, e.g., drop-out on fully connected layers, batch normalization in convolution layers and data augmentation techniques were applied. Given that, a CNN model having more layers and being pre-trained with large datasets (e.g., ImageNet) might have better feature representation capacity, we also performed our analysis by using ResNet50 [53], which gave significantly better (p-value < 0.01) VAD results as compared to AlexNet. This is performed as follows.

When multiple sub-dynamic images, each representing a single person at a time and each has a VAD label either speaking or not-speaking, are the inputs, ResNet50 (pre-trained on ImageNet dataset) is fine-tuned in two different ways:

i) Fine-tuning 1 (FT1): We include an additional fully connected layer (shown as $Fc1$ in Figure 2) after the final fully connected layer of ResNet50 (shown as Fc in Figure 2). $Fc1$ layer has 2048 neurons and its weights are randomly initialized. During this fine-tuning, only the weights of Fc and $Fc1$ are updated (implying that convolutional layers, e.g., Block 4 and 5 in Figure 2 are not updated) when the learning rate is equal to 10^{-5}. After this fine-tuning, $Fc1$ features are extracted.

ii) Fine-tuning 2 (FT2): The convolutional layers shown as Block 4, Block 5 and the fully connected layer Fc in Figure 2 are updated. The weights of Block 4, 5 and Fc are randomly initialized and training is performed with $5{\times}10^{-6}$ learning rate. After this fine-tuning, Fc features are extracted.

We have not trained the earlier ResNet50 layers than Block 4, to utilize the information already learned (e.g., edges, corners and very general image features) by pre-training ResNet50. This is based on the assumption that such features can be common even in different image domains (in our case ImageNet data and dynamic images).

FT1 and FT2 are applied with an end-to-end manner with cross entropy loss function, and Adam optimizer for 20 (for FT1) and 10 epochs (for FT2). The parameter values, e.g., learning rate and the number of epochs were determined based on the VAD performance obtained on the validation set. Validation set is 10% of the total training set in a cross-validation fold. It is randomly selected and not overlapping with the training set used in the fine-tuning. At each batch, 128 randomly selected dynamic images (64 speaking and 64 not-speaking) were used. It is important to balance the number of speaking and not-speaking dynamic images in each batch as the used datasets have much more not-speaking bounding boxes as compared to the speaking bounding boxes, and training with an imbalanced data misleads the classification task a lot. Additionally, data augmentation is applied as follows. Some randomly selected training dynamic images are horizontally flipped and/or a $64{\times}64$ randomly selected patch is replaced with the mean value of the images, which can be seen as a dropout in input layer.

C. Unsupervised Domain Adaptation

The unsupervised domain adaption contains a dual-layer stacked sparse autoencoder (2AE) and a PCA-based subspace alignment technique [55] (SA), which are applied together with SVM classifier learning.

The autoencoder is essentially a neural network trained to map the input to an output, which is a reconstruction of the input. The parameters of stacked sparse autoencoder: the number of hidden neurons in the first and second layers, $l2$ weight regularization, sparsity regularization and sparsity proportion are determined based on the best VAD performance of SVM applied to the validation set.

The set of values used as the number of hidden neurons in the first layer is $\{1024, 512, 256, 128\}$, the number of hidden neurons in the second layer is $\{64, 128, 256, 512\}$, the sparsity regularization is $\{1, 2, 4, 8, 16, 32\}$, $l2$ weight regularization is $\{0.00002, 0.0002, 0.002, 0.02, 0.2\}$ and the sparsity proportion

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3007350, IEEE Transactions on Multimedia

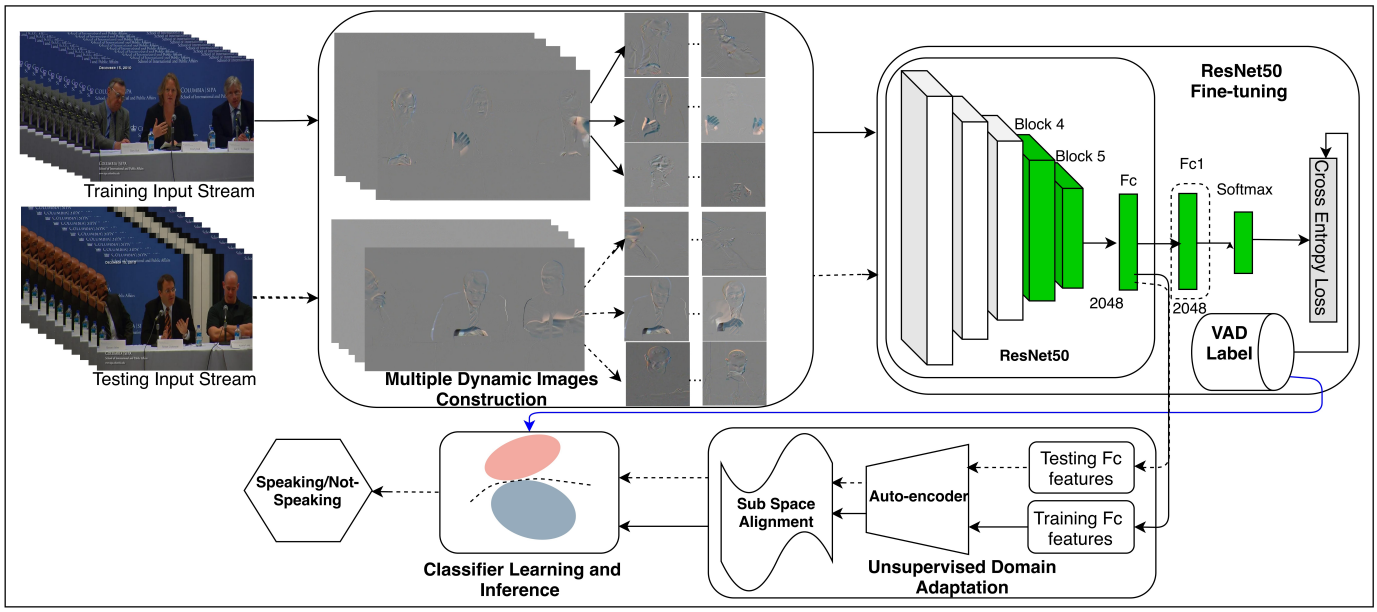

Fig. 2: The illustration of the proposed methodology. It includes multiple dynamic images construction, ResNet50 training to obtain visual feature representations and unsupervised domain adaptation, which is jointly applied with a classifier learning. Block 4 and 5 are convolution layers, Fc and $Fc1$ are fully connected layers. Block 4, 5 and Fc already exist in the ResNet50 architecture while we have included $Fc1$ as an additional layer. Visual feature representations are extracted from either Fc or $Fc1$ layer depending on the fine-tuning strategy followed. Herein, both layers are shown but as an example, Fc features are given to the unsupervised domain adaptation component.

is $\{0.005, 0.05, 0.1, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$. The autoencoder training is performed with a balanced way, which means using the complete test set (target domain) and the same amount of randomly selected training set (source domain) for e number of epochs while e=$\{100, 200, 300, 400\}$.

Given a stacked sparse autoencoder (2AE) trained in a balanced way with a set of parameters, the new feature representations out of it are obtained not only for the data it is trained with, but for the complete training (Tr'), validation (V') and test (T') sets. The PCA-based subspace alignment mechanism (SA) [55] is performed using Tr', V' and T' and it results in features $Tr'_{aligned}$, $V'_{aligned}$ and $T'_{aligned}$, respectively. $Tr'_{aligned}$ is used to learn an SVM classifier, which is later used to infer the VAD labels of $V'_{aligned}$ (utilized to determine the best performing set of parameters) and $T'_{aligned}$ (corresponds to the final prediction of the proposed method).

Since using autoencoder (2AE) does not always guarantee effective features, PCA-based subspace alignment mechanism (SA) [55] is applied after autoencoder. The main objective of SA is to bring the full sets of training and test (validation) in a common space. The only parameter of SA is the number of eigenvectors, whose value was determined from a set of values: $\{5, 10, 15, 20, 25, 30\}$ based on the best VAD performance of SVM applied to the validation set.

D. Classifier Learning and Inference

A linear SVM, with the regularization parameter C taken as 10^k while $k = \{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$ is trained together with the aforementioned unsupervised domain adaptation method (2AE+SA). The trained classifier is used for the classification of validation and test sets.

V. Experimental Analyses

This section includes the experimental analyses applied using publicly available real-world datasets: Columbia [30] and RealVAD as well as the cross-dataset experiments. To train and test the proposed method, we used the bounding boxes provided by the RealVAD dataset. For Columbia dataset, as the supplied bounding boxes include only the head and shoulders of the subjects, i.e., not the whole upper body, we re-extracted the bounding boxes accordingly (see Figure 2 sub-dynamic images to infer the coverage of the bounding boxes).

A. Experiments on Columbia Dataset [30]

Columbia dataset [30] is constituted from an 87 minutes-long YouTube video that is from a panel discussion at Columbia University. In that video, there are in total seven panelists. The field of the view of the camera is changing to focus on smaller groups of panelists at a time. To perform a fair comparison, following the baseline works [10], [30], [35], [54], we only focused on the parts of the video where there is more than one person in the frame and discarded any person in the margins of the video. This resulted in five panelists (Bell, Bollinger, Lieberman, Long, Sick) while two/three of them are visible in a video frame.

Columbia dataset contains the visual data annotations: the bounding boxes (each covers the head and shoulders of a single panelist) and the associated VAD labels. This annotation corresponds to 35 minutes of the YouTube video. We used the

VAD labels as it was provided and extracted the bounding boxes containing the whole upper body (from head to waist including arms and hands when they are visible) of each panelist.

Following the SOA methods: [10], [30], [35], leave-one-panelist-out cross validation (in each fold of cross-validation, training set is composed of the data belonging to four panelists while test includes the data belonging to one remaining panelist) and F1-score as the evaluation metric were used to make performance comparisons between the proposed method and them.

Using this dataset, an ablation study for the proposed method was also performed. The ablation study allowed us to compare the performance of:
i) AlexNet (DI+AlexNet+Softmax) and ResNet50 (DI+FT1+Softmax and DI+FT2+Softmax),
ii) ResNet50 w/ and w/o data augmentation,
iii) One-layer sparse autoencoder (DI+FT1+AE+SVM) and two-layers stacked sparse autoencoder (DI+FT1+2AE+SVM),
iv) Sparse autoencoder w/ and w/o balanced training,
v) The proposed method (DI+FT1+2AE+SA+SVM and DI+FT2+2AE+SA+SVM) and the proposed method without sparse autoencoder, i.e. using subspace alignment only for the unsupervised domain adaptation (DI+FT1+SA+SVM and DI+FT2+SA+SVM),
vi) The proposed method (DI+FT1+2AE+SA+SVM and DI+FT2+2AE+SA+SVM) and the proposed method without subspace alignment, i.e. using sparse autoencoder only for the unsupervised domain adaptation (DI+FT1+2AE+SVM and DI+FT2+2AE+SVM),
vii) FT1 and FT2 for each setting, i.e., DI+FT1+Softmax, DI+FT2+Softmax, DI+FT1+SVM, DI+FT2+SVM, DI+FT1+2AE+SVM, DI+FT2+2AE+SVM, DI+FT1+SA+SVM, DI+FT2+SA+SVM, DI+FT1+2AE+SA+SVM and DI+FT2+2AE+SA+SVM.

All the corresponding results are given in Table III. As seen, the proposed method with FT2 (DI+FT2+2AE+SA+SVM) performs better than the proposed method with FT1 (DI+FT1+2AE+SA+SVM). At the same time, DI+FT2+2AE+SA+SVM outperforms all the baseline methods on average (93.36%), and for three panelists out of five, it performs the best of all methods. The performances of [10], [30], [35] are highly dependent to the choice of window size (W) of temporal continuity algorithm, which post-processes the VAD predictions considering the predictions of consecutive video frames. In detail, it assumes that if a person is speaking it is more likely that she/he will continue speaking for a while rather than stop speaking. Given that we create dynamic images for each 10 consecutive frames, it can be more fair to compare the performance of the proposed method when W is equal to 10. Nevertheless, DI+FT1+2AE+SA+SVM and DI+FT2+2AE+SA+SVM perform better than [30], [35] for any W values used in those studies.

Furthermore, DI+FT2+2AE+SA+SVM performs better than [10], which is a multimodal VAD approach. The performance of DI+FT2+2AE+SA+SVM is a very important achievement given that it is based on only one modality, i.e., the upper body motion. The result of DI+FT2+Softmax (89.66%) proves the superiority of the features extracted such that it is better than the majority of the methods and particularly better than the features extracted from FT1 (86.24%).

In addition to the better average VAD performance of the proposed method with FT2, the fact that it has low VAD standard deviation (STD) while still performing the best, is also a significant aspect. In detail, the performances of baselines, e.g., [30], [35] have fluctuations such that they perform well for some persons (e.g., Long: 86.90%) but their performances are not generalizable. On the other hand, the performance of the proposed method is much consistent (STD= 3.23), showing the importance of applying unsupervised domain adaptation with better features.

The results of ablation study in Table III show that, FT1 and FT2 of ResNet50 significantly outperform fine-tuning AlexNet. Data augmentation applied in ResNet50 fine-tuning (DI+FT1+Softmax) improves the VAD performance for all subjects. Balanced training of sparse autoencoder contributes positively to the VAD performances for all subjects independent to the number of layers (one-layer: DI+FT1+AE+SVM) or two-layers stacked sparse autoencoder: DI+FT2+AE+SVM). Moreover, two-layers stacked autoencoder (DI+FT1+2AE+SVM) enhances the VAD performance, which also results in lower-dimensional feature space as compared to the single layer autoencoder (DI+FT1+AE+SVM). Using SVM (DI+FT1+SVM and DI+FT2+SVM) performs slightly better than using Softmax (DI+FT1+Softmax and DI+FT2+Softmax). Applying SA (DI+FT1+SA+SVM and DI+FT2+SA+SVM) additionally, brings small improvements compared to the performance of SVM alone (DI+FT1+SVM and DI+FT2+SVM). DI+FT1+2AE+SVM, on the other hand, performs better than DI+FT1+SVM while for DI+FT2+2AE+SVM, this is not observed. Applying 2AE together with SA (i.e., DI+FT1+2AE+SA+SVM and DI+FT2+2AE+SA+SVM) performs better than all, showing the importance of using these two components together.

B. Experiments on RealVAD Dataset

We benchmarked the RealVAD dataset by performing the following experimental analyses.
i) The proposed method with FT1 (DI+FT1+2AE+SA+SVM).
ii) The proposed method with FT2 (DI+FT2+2AE+SA+SVM).
iii) Applying only the multiple dynamic image construction and FT1, i.e., not applying unsupervised domain adaptation and SVM, but instead using Softmax as the classifier (DI+FT1+Softmax).
iv) Applying only the multiple dynamic image construction and FT2, i.e., not applying unsupervised domain adaptation and SVM, but instead using Softmax as the classifier (DI+FT2+Softmax).

All of these analyses were performed with leave-one-panelist-out setting such that at each fold of cross-validation, training set is composed of the data belonging to eight panelists while test includes the data belonging to the one remaining panelist. As mentioned before, to learn the optimum

TABLE III: F1-scores on the Columbia dataset [30]. AVG and STD stand for the average and standard deviation of F1-scores of all speakers, respectively. W stands for the window size (see text and the corresponding reference for more information). DI, FT, w/AUG, AE, 2AE, BT, SA, SOA stand for dynamic image, fine-tuning, with augmentation, sparse autoencoder, stacked sparse autoencoder, balanced training, subspace alignment, and state of the art, respectively. The best result of each column is emphasized in bold-face.

Method	Bell	Bollinger	Lieberman	Long	Sick	AVG	STD	Details
Ablation Study								
DI+AlexNet+Softmax	78.29	84.38	59.39	63.59	64.14	69.96	10.76	FT AlexNet
DI+FT1+Softmax	85.95	91.08	90.71	71.84	85.95	85.11	7.82	FT1 ResNet50
DI+FT1+Softmax	86.07	93.30	91.88	73.62	86.34	86.24	7.76	FT1 ResNet50 w/AUG
DI+FT2+Softmax	89.06	95.44	92.11	82.76	88.95	89.66	4.69	FT2 ResNet50 w/AUG
DI+FT1+SVM	86.35	93.78	92.34	76.09	86.25	86.96	6.97	FT1 ResNet50 w/AUG, Fc1 features.
DI+FT2+SVM	90.39	96.67	91.58	87.30	89.77	91.14	3.46	FT2 ResNet50 w/AUG, Fc features.
DI+FT1+AE+SVM	86.54	92.95	92.19	77.57	86.96	87.24	6.15	FT1 ResNet50 w/AUG, Fc1 features, 1-layer AE.
DI+FT1+AE+SVM	87.18	93.58	92.10	78.22	87.39	87.69	6.00	FT1 ResNet50 w/AUG, Fc1 features, BT 1-layer AE.
DI+FT1+2AE+SVM	87.28	94.01	92.20	80.70	87.35	88.31	5.18	FT1 ResNet50 w/AUG, Fc1 features, BT 2-layers AE.
DI+FT2+2AE+SVM	89.77	96.45	92.13	85.64	89.52	90.70	3.55	FT2 ResNet50 w/AUG, Fc features, BT 2-layers AE.
DI+FT1+SA+SVM	86.51	94.12	92.33	77.32	86.87	87.43	6.56	FT1 ResNet50 w/AUG, Fc1 features, PCA-SA.
DI+FT2+SA+SVM	90.34	98.30	93.48	86.07	91.53	91.94	4.00	FT2 ResNet50 w/AUG, Fc features, PCA-SA.
Proposed Method								
DI+FT1+2AE+SA+SVM	87.28	96.35	92.15	83.03	87.21	89.20	5.14	FT1 ResNet50 w/AUG, Fc1 features, BT 2-layers AE, PCA-SA
DI+FT2+2AE+SA+SVM	91.92	**98.90**	**94.05**	89.07	**92.84**	**93.36**	**3.23**	FT2 ResNet50 w/AUG, Fc features, BT 2-layers AE, PCA-SA
SOA								
[30], [35]	82.90	65.80	73.60	86.90	81.80	78.20	8.45	W=10, [30], [35] using head and shoulders motion.
[30], [35]	90.30	69.00	82.40	96.00	89.30	85.40	10.36	W=100, [30], [35] using head and shoulders motion.
[10]	**93.70**	83.40	86.80	**97.70**	86.10	89.54	5.94	W=10, [10] using audio and lips motion.

values of parameters, we used a randomly chosen validation set composed of equal amount of speaking and not-speaking bounding boxes, in total equal to 10% of the training set while not intersecting with the data used for training.

The corresponding results are given in Table IV in terms of true positive (TP), true negative (TN), false positive (FP), false negative (FN) rates and F1-score for each panelist. The results show that, out of all methods, the best average TP score belongs to DI+FT1+2AE+SA+SVM (88.41%), the best average TN score belongs to DI+FT1+Softmax (70.05%) and the best average F1-score belongs to DI+FT2+2AE+SA+SVM (53.04%). The F1-scores DI+FT1+2AE+SA+SVM and DI+FT2+2AE+SA+SVM are almost the same. On the other hand, the average TP score of DI+FT2+Softmax (84.11%) is significantly better (p-value < 0.01) than the average TP score DI+FT1+Softmax (74.81%) while F1-score of these two methods are similar (49.54% and 49.17%, respectively). The average TP, TN and F1-scores of DI+FT2+Softmax are approx. 4% less than DI+FT2+2AE+SA+SVM (TP=87.44%, TN=68.17%, F1-score=53.04%). In other words, the unsupervised domain adaptation with SVM increases the average TP, TN and F1-scores for approx. 4%. The worst F1-scores of DI+FT1+2AE+SA+SVM and DI+FT2+2AE+SA+SVM (proposed method) are obtained for the panelist three (P3). The worst TP scores of them are obtained for the panelists two (P2) and three (P3). The TN scores of the proposed method for P2 and P3 are as well as other TN scores.

C. Cross-Dataset Analyses

The cross-dataset analyses performed can be described as follows:

i) The whole Columbia dataset [30] was used as the training set and a single panelist of the RealVAD dataset was used as the test set when DI+FT2+2AE+SA+SVM was applied. Additionally, to better understand the effectiveness of the domain adaptation component (2AE+SA+SVM), we trained DI+FT2+Softmax using the whole Columbia dataset and tested it on the RealVAD dataset. The corresponding results are given in Table V.

ii) Similarly, the whole RealVAD dataset was used as the training set and a single panelist of the Columbia dataset was used as the test set when DI+FT2+2AE+SA+SVM was applied. We also trained DI+FT2+Softmax using the whole RealVAD dataset and tested it on the Columbia dataset. The corresponding results can be found in Table VI.

Given the overall better performance of FT2 compared to FT1, all the experimental analyses given in this section were realized using FT2. Additionally, we reported the results of random guess. Random guess was computed by randomly generating a VAD label for each observation in the test set and then calculating the scores (true positive (TP), true negative (TN), false positive (FP), false negative (FN) rates and F1-score) based on them, for 1000 individual times. The reported scores are the average of these 1000 scores.

The average F1-score of the proposed method when cross-dataset analysis is applied (DI+FT2+2AE+SA+SVM=51.52%, Table V) is worse than the average F1-score of the proposed method trained and tested on the same dataset (DI+FT2+2AE+SA+SVM=53.04%, Table IV). Particularly, the average TN rate of cross-dataset experiment is lower (64.75% vs. 68.17%). Similarly, the average F1-score of DI+FT2+Softmax when cross-dataset analysis is applied (41.04%, Table V) is worse than the average F1-score of DI+FT2+Softmax trained and tested on the same dataset (49.54%, Table IV). The average TP rate of DI+FT2+Softmax significantly decreases when cross-dataset experiment is applied (from 84.11% to 62.39%). The performance decline observed in Table V as compared to the results in Table IV

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3007350, IEEE Transactions on Multimedia

TABLE IV: The VAD performances in terms of true positive (TP), true negative (TN), false positive (FP), false negative (FN) rates and F1-score for each panelist when the experimental analyses given in text were applied using RealVAD dataset. Positive class refers to speaking while negative class represents not-speaking. AVG and STD stand for average and standard deviation of each metric for all panelists, respectively. DI, FT, w/AUG, 2AE, and SA stand for dynamic image, fine-tuning, with augmentation, two-layer stacked sparse autoencoder, and subspace alignment, respectively. The best results are emphasized in bold-face.

DI+FT1+2AE+SA+SVM											
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**AVG**	**STD**
TP (%)	91.27	**75.49**	70.78	94.96	**89.17**	**98.20**	**94.35**	**89.20**	**92.27**	**88.41**	9.19
TN (%)	68.50	70.95	**68.87**	**75.74**	61.43	61.71	56.98	68.59	70.20	67.00	5.81
FP (%)	31.50	29.05	**31.13**	**24.26**	38.57	38.29	43.02	31.41	29.80	33.00	5.81
FN (%)	8.73	**24.51**	29.22	5.04	**10.83**	**1.80**	**5.65**	**10.80**	**7.73**	**11.59**	9.19
F1-score (%)	**53.88**	**53.94**	**43.64**	**53.58**	**44.70**	47.17	58.10	65.07	54.36	52.72	6.76
DI+FT2+2AE+SA+SVM											
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**AVG**	**STD**
TP (%)	**96.79**	72.30	**71.72**	**97.99**	85.08	90.69	**94.22**	87.08	91.05	87.44	9.68
TN (%)	61.95	**72.72**	67.07	72.47	63.73	**70.41**	**58.23**	**74.39**	**72.56**	68.17	5.71
FP (%)	38.05	**27.28**	32.93	27.53	36.27	**29.59**	**41.77**	**25.61**	**27.44**	31.83	5.71
FN (%)	**3.21**	27.70	**28.28**	**2.01**	14.92	9.31	**5.78**	12.92	8.95	12.56	9.69
F1-score (%)	51.63	53.49	42.92	51.70	44.40	**50.48**	58.73	**67.94**	55.75	**53.04**	7.5
DI+FT1+Softmax											
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**AVG**	**STD**
TP (%)	72.79	65.69	43.59	93.98	75.09	77.27	80.65	76.95	87.25	74.81	14.27
TN (%)	69.55	71.77	73.00	**75.41**	**67.78**	56.79	69.08	71.02	75.67	**70.05**	5.68
FP (%)	30.45	28.23	26.6	24.59	**32.22**	43.21	30.92	28.98	24.33	**29.95**	5.68
FN (%)	27.21	34.31	56.41	6.02	24.91	22.73	19.35	23.05	12.75	25.19	14.27
F1-score (%)	46.15	49.18	31.72	52.84	42.77	44.03	**58.79**	60.27	**56.75**	49.17	9.14
DI+FT2+Softmax											
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**AVG**	**STD**
TP (%)	77.02	70.83	55.78	95.7	85.56	99.25	91.08	**89.87**	91.86	84.11	13.86
TN (%)	**71.66**	72.86	68.63	74.41	61.12	54.33	57.07	53.68	72.07	65.09	8.50
FP (%)	**28.34**	27.14	31.37	25.59	38.88	45.67	42.93	46.32	27.93	34.91	8.50
FN (%)	22.98	29.17	44.22	4.30	14.44	0.75	8.92	**10.13**	8.14	15.89	13.86
F1-score (%)	49.73	52.79	35.81	52.58	43.00	43.20	56.71	57.03	55.02	49.54	7.32

TABLE V: The results of cross-dataset analyses: the Columbia dataset [30] was used for training while the RealVAD dataset was the test set. The evaluation metrics are true positive (TP), true negative (TN), false positive (FP), false negative (FN) rates and F1-score. Positive class refers to speaking while negative class represents not-speaking. AVG and STD stand for the average and standard deviation of each metric for all panelists, respectively. The best results are shown in bold-face.

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**AVG**	**STD**
Cross-dataset Analysis: DI+FT2+Softmax											
TP (%)	73.43	49.27	33.60	54.87	53.67	74.63	76.88	63.92	81.28	62.39	15.71
TN (%)	49.65	**77.94**	**74.42**	**75.47**	**73.44**	**65.74**	**62.33**	65.91	71.82	**68.52**	8.77
FP (%)	50.35	**22.06**	**25.58**	**24.53**	**26.56**	**34.26**	**37.67**	34.09	28.18	**31.48**	8.77
FN (%)	26.57	50.73	66.40	45.13	46.33	25.37	23.12	36.08	18.72	37.61	15.71
F1-score (%)	35.88	43.11	25.86	34.71	35.92	40.36	52.83	49.92	50.72	41.04	8.94
Cross-dataset Analysis: DI+FT2+2AE+SA+SVM											
TP (%)	**94.22**	**85.66**	**66.88**	**98.14**	**89.29**	**99.55**	**97.24**	**92.81**	**73.83**	**88.62**	11.37
TN (%)	**66.34**	58.34	67.86	70.70	46.98	**65.59**	52.01	**68.91**	**86.04**	64.75	11.39
FP (%)	**33.66**	41.66	32.14	29.30	53.02	**34.41**	47.99	**31.09**	**13.96**	35.25	11.39
FN (%)	**5.78**	**14.34**	**33.13**	**1.86**	**10.71**	**0.45**	**2.76**	**7.19**	**26.17**	**11.38**	11.37
F1-score (%)	**53.56**	**51.08**	**41.09**	**50.22**	**37.29**	**50.32**	**56.74**	**53.58**	**69.79**	**51.52**	9.25
Random Guess											
TP (%)	50.00	49.99	49.98	50.02	50.03	50.00	49.97	50.00	50.02	50.00	0.02
TN (%)	50.00	49.99	50.01	50.00	49.99	50.00	50.00	50.00	49.99	50.00	0.01
FP (%)	50.00	50.01	49.99	50.00	50.01	50.00	50.00	50.00	50.01	50.00	0.01
FN (%)	50.00	50.01	50.02	49.98	49.98	50.00	50.03	50.00	49.98	50.00	0.02
F1-score (%)	25.88	30.31	25.20	20.89	23.64	23.01	33.17	35.05	25.21	26.93	4.82

TABLE VI: The results of cross-dataset analyses: The RealVAD was used for training while the Columbia dataset [30] was the test set. The evaluation metrics are true positive (TP), true negative (TN), false positive (FP), false negative (FN) rates and F1-score obtained for each panelist. Positive class refers to speaking while negative class represents not-speaking. AVG and STD stand for the average and standard deviation of each metric for all panelists, respectively. The best results are shown in bold-face.

	Bell	**Boll.**	**Sick**	**Long**	**Lib.**	**AVG**	**STD**
Cross-dataset Analysis: DI+FT2+Softmax							
TP (%)	92.71	96.73	**99.31**	**98.10**	**97.49**	**96.87**	2.51
TN (%)	54.65	75.04	40.01	52.47	53.85	55.20	12.59
FP (%)	45.35	24.96	59.99	47.53	46.15	44.80	12.59
FN (%)	7.29	3.27	**0.69**	**1.91**	**2.51**	**3.13**	2.51
F1-score (%)	77.30	86.62	64.84	53.08	82.63	72.89	13.78
Cross-dataset Analysis: DI+FT2+2AE+SA+SVM							
TP (%)	**96.57**	**98.56**	94.41	95.82	93.85	95.84	**1.87**
TN (%)	**70.75**	**79.45**	**72.20**	**72.63**	**77.91**	**74.59**	**3.84**
FP (%)	**29.25**	**20.55**	**27.80**	**27.37**	**22.09**	**25.41**	**3.84**
FN (%)	**3.43**	**1.44**	5.59	4.18	6.15	4.16	**1.87**
F1-score (%)	**84.80**	**89.63**	**77.39**	**65.06**	**88.43**	**81.06**	**10.14**
Random Guess							
TP (%)	49.98	49.97	49.99	50.02	50.01	49.99	0.02
TN (%)	50.00	50.00	50.00	50.00	50.01	50.00	0.01
FP (%)	50.00	50.00	50.00	50.00	49.99	50.00	0.01
FN (%)	50.02	50.03	50.01	49.98	49.99	50.01	0.02
F1-score (%)	49.49	49.16	41.80	30.29	52.18	44.58	8.87

are expected as it includes two different datasets not having much in common except being based on panel discussions.

On the other hand, the scores of DI+FT2+2AE+SA+SVM as compared to the scores of DI+FT2+Softmax in Table V clearly show that 2AE+SA+SVM positively contributes to the VAD performance a lot (e.g., 10.48% improvement on average F1-score) by handling the domain-shift problem among different datasets. DI+FT2+2AE+SA+SVM and DI+FT2+Softmax both perform significantly better than the random guess for all metrics.

When cross-dataset analysis is applied, the average F1-scores of the proposed method (DI+FT2+2AE+SA+SVM=81.06%, Table VI) as well as DI+FT2+Softmax (72.89%, Table VI) are worse than the average F1-scores of the proposed method (DI+FT2+2AE+SA+SVM=93.36%, Table III) and DI+FT2+Softmax (89.66%, Table III) trained and tested on the same dataset. 2AE+SA+SVM once again positively contributes to the VAD performance, showing 8.17% improvement compared to DI+FT2+Softmax on average F1-score (Table VI). DI+FT2+2AE+SA+SVM and DI+FT2+Softmax both result in performances significantly better than the random guess for all metrics.

The average F1-scores (51.52% and 41.04%) obtained for RealVAD dataset are much worse than the average F1-scores (81.06% and 72.89%) obtained for Columbia dataset [30], which can be interpreted as our dataset is more challenging.

D. Proposed Method with a Real-time Manner

The experiments with unsupervised domain adaptation (2AE+SA+SVM) in Tables III, IV, V and VI were performed when all dynamic images belonging to one person were used as the target set. However, it is possible to apply the proposed method with a real-time manner, which was realized as described below and the corresponding results are given in Table VII in terms of F1-score.

TABLE VII: F1-scores on the Columbia dataset [30] when the proposed method is applied in a real-time manner: DI+FT2+2AE+SA+SVM_realTime (see text for details).

	Bell	**Boll.**	**Sick**	**Long**	**Lie.**	**AVG**	**STD**
F1-score (%)	89.52	96.86	88.97	84.41	93.10	90.57	4.68

We trained 2AE+SA+SVM component of the proposed method with a target set composed of one dynamic image and the source set containing all dynamic images belonging to persons different from the one in the target set. As can be observed, balanced training of 2AE is not feasible. The learned SVM model was used to classify only that target set. In other words, for each test dynamic image, we modelled a new 2AE+SA+SVM. This procedure was tested on Columbia dataset [30] when ResNet50 training and feature extraction part of the proposed method was realized with FT2.

The average F1-score obtained (90.57%) when the proposed method is applied with a real-time manner (DI+FT2+2AE+SA+SVM_realTime) is worse than the average F1-score of the proposed method (DI+FT2+2AE+SA+SVM 93.36%, Table III). This means that the effectiveness of unsupervised domain adaptation in DI+FT2+2AE+SA+SVM_realTime is fewer. This is expected as the size of source set is a lot bigger than the size of target set in 2AE training. On the other hand, DI+FT2+2AE+SA+SVM_realTime still performs better than all SOA methods: [10], [30], [35] on average as well as for three panelists (see Table III).

VI. Discussions

Voice activity detection (VAD) is traditionally performed by audio processing. However, the use of visual cues instead of audio cues, namely visual VAD (VVAD), can be important especially in case audio is neither feasible to acquire nor reliable. The existing VVAD studies relied on the detection of individual body parts, e.g., lips [10], [13], [14], face [11], [13], [14], [26], head [13], [14], [16], gaze [16], which might not be always very effective. For instance, if the speakers are sitting in a single row, it is rare that they face each other, which makes gaze-based methods fail. Detecting lips motion is not always possible, e.g., when speaker presents a profile view to the camera or the camera resolution is low, or the speaker is far away from the camera or her/his hands occlude the speaker's face.

In this study, we have presented a novel vision-based VAD method. It is based on holistic motion analysis, thus, unlike previous studies, it does not require detection of specific body parts. We have utilized dynamic image representation [52] with an end-to-end deep learning-based methodology to extract novel features from upper body. Psychology literature confirms that the way people move while speaking varies a lot from person to person (e.g., one can be moving/ gesticulating more than others). This might result in a domain-shift problem, i.e.,

dealing with unrelated distributions of body motion representations in the training and test data, which leads to performance degradation. We have taken on this challenge by integrating a simple but effective domain adaptation technique.

The proposed method performed not only better than the SOA VVAD methods, but also better than the SOA multimodal VAD approach when a real-world dataset was used. The ablation study proved the superiority of the proposed method and demonstrated the positive contribution of each component. The proposed method is a generic approach, its ResNet50 fine-tuning and feature extraction component supports person-invariant training such that it can be applied to a subject who has never been seen by the CNN model. Furthermore, the proposed method provides more consistent VAD results; such that the detection performances are equally well for all persons.

A limitation of the proposed method is requiring the number of raw video frames to construct one single dynamic image, which is fixed to 10 in this study. The other parameters are found automatically based on the best VAD performance obtained on the validation set. As future work, the proposed method will be adapted to work on data streams having more camera motion such as egocentric videos.

The VAD datasets in the literature have played a catalyst role to breakthrough many approaches. However, most of these datasets are limited to address many real-world challenges. They usually include short and relatively simple scenarios. Majority of them are based on role-plays, which were captured in a lab setting where the subjects do not behave naturally as they perform the imposed role instead of a spontaneous behavior. Motivating from this, as a major contribution of this paper, we have introduced a new multimodal real-world VAD dataset called RealVAD.

One main characteristics of RealVAD dataset, which to the best of our knowledge, have not been specifically considered during collection of any other VAD dataset, is containing subjects with different ethnicities. This can be particularly critical to evaluate VVAD methods, as it is highly possible that people from different cultures behave/gesticulate differently. Furthermore, the RealVAD dataset includes many other real-world challenges/situations, some of those have not been addressed in other datasets. It contains nine panelists at a time, which is the highest number of persons among real-world datasets. Its VAD ground-truth was constructed by human annotators while only the labels having perfect agreement were used. We believe that the way we obtained the ground-truth is less erroneous as compared to using an automatic speaker diarization algorithm to obtain voice activity as applied in, e.g., [15], [30], [45], [46]. RealVAD contains imbalanced VAD classes such that the total number of not-speaking detections is more than the total number of speaking detections. This imbalance introduces a new challenge to the targeted classification problem. Moreover, some of the frames have background motions and occlusions, which were intentionally avoided challenges during the creation of many other datasets.

One limitation of RealVAD dataset is not containing simultaneous speakers. Given that, our method is applied to every single subject in a given frame independently, having (or not having) simultaneous speakers does change its performance. However, we agree that having simultaneous speakers would result in a better benchmark especially to test audio only or audiovisual VAD approaches.

The proposed method was also applied to RealVAD dataset and in that way we have benchmarked our dataset. The results of the cross-dataset analyses showed that RealVAD dataset is more challenging as compared to the only other publicly available real-world dataset [30]. Furthermore, the cross-dataset analyses proved the positive contribution of the unsupervised domain adaptation component as applying it improved the VAD performance remarkably.

In the future, our dataset can be used to evaluate novel VAD methodologies, which are not necessarily solely based on visual data processing but also can be audio-based or multimodal. The investigation of domain adaptation in the context of VAD is a very recent issue. The proposed dataset can contribute to examine this deeply, such as to study the challenges brought by ethnic differences to develop robust VAD technologies. Finally, another valuable characteristic of RealVAD dataset is that it can be used for evaluating different applications, e.g., active speaker localization, speaker recognition and nonverbal behavior analysis after collecting appropriate annotations.

REFERENCES

[1] M. Cristani, A. Pesarin, A. Vinciarelli, M. Crocco, and V. Murino, "Look at who's talking: Voice activity detection by automated gesture analysis," in *Proceedings of International Joint Conference on Ambient Intelligence*, 2011, pp. 72–80.

[2] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1086–1099, 2018.

[3] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 74–79.

[4] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung, "The matchnmingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates," *IEEE Transactions on Affective Computing*, 2018.

[5] 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.

[6] A. Abushakra and M. Faezipour, "Acoustic signal classification of breathing movements to virtually aid breath regulation," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 2, pp. 493–500, March 2013.

[7] Y. Hu, J. Kautz, Y. Yu, and W. Wang, "Speaker-following video subtitles," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 11, no. 2, p. 32, 2015.

[8] B. G. Gebre, P. Wittenburg, and T. Heskes, "The gesturer is the speaker," in *Proceedings of IEEE ICASSP*, 2013, pp. 3751–3755.

[9] H. Vajaria, S. Sarkar, and R. Kasturi, "Exploring co-occurence between speech and body movement for audio-guided video localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1608–1617, 2008.

[10] J. S. Chung and A. Zisserman, "Learning to lip read words by watching videos," *Computer Vision and Image Understanding*, vol. 173, pp. 76–85, 2018.

[11] K. Stefanov, J. Beskow, and G. Salvi, "Vision-based active speaker detection in multiparty interaction," in *Int. Workshop Grounding Language Understanding*, 2017, pp. 47–51.

[12] J. M. Rehg, K. P. Murphy, and P. W. Fieguth, "Vision-based speaker detection using bayesian networks," in *Proceedings of IEEE CVPR*, vol. 2, 1999, pp. 110–116.

[13] B. Joosten, E. Postma, and E. Krahmer, "Voice activity detection based on facial movement," *Journal on Multimodal User Interfaces*, vol. 9, no. 3, pp. 183–193, 2015.

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

[15] P. Chakravarty, S. Mirzaei, T. Tuytelaars, and H. V. hamme, "Who's speaking?: Audio-supervised classification of active speakers in video," in *Proceedings of ACM ICMI*, 2015, pp. 87–90.

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

[17] N. Latif, A. V. Barbosa, E. Vatiokiotis-Bateson, M. S. Castelhano, and K. G. Munhall, "Movement coordination during conversation," *PLOS ONE*, vol. 9, no. 8, pp. 1–10, 2014.

[18] P. Feyereisen and J.-D. de Lannoy, "Gestures and speech: Psychological investigations," *Cambridge University Press*, 1991.

[19] D. McNeill, "So you think gestures are nonverbal," *Psychological review*, vol. 92, no. 3, pp. 350–350, 1985.

[20] N. Campbell and N. Suzuki, "Working with very sparse data to detect speaker and listener participation in a meetings corpus," in *Workshop Programme*, 2006.

[21] S. Kopp and I. Wachsmuth, "Synthesizing multimodal utterances for conversational agent," *Computer Animation and Virtual Worlds*, vol. 15, no. 1, pp. 39–52, 2004.

[22] C. Beyan, M. Shahid, and V. Murino, "Investigation of small group social interactions using deep visual activity-based nonverbal features," in *Proceedings of ACMMM*, 2018, pp. 311–319.

[23] "Privacy camp 2018: Round-table, government hacking in different national contexts and strategies." https://www.youtube.com/watch?v=51pRTOIso4U, accessed: 9 December 2019.

[24] F. Tao, J. H. L. Hansen, and C. Busso, "An unsupervised visual-only voice activity detection approach using temporal orofacial features," in *INTERSPEECH*, 2015.

[25] A. Krizhevsky, S. Ilya, and H. Geoffrey, "Imagenet classification with deep convolutional neural networks," in *Proceedings of Advances in Neural Information Processing Systems (NeuroIPS)*, 2012, pp. 1097–1105.

[26] K. Stefanov, J. Beskow, and G. Salvi, "Self-supervised vision-based detection of the active speaker as support for socially-aware language acquisition," *IEEE Transactions on Cognitive and Developmental Systems*, 2019.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[28] B. G. Gebre, P. Wittenburg, T. Heskes, and S. Drude, "Motion history images for online speaker/signer diarization," in *Proceedings of (ICASSP)*, 2014, pp. 1537–1541.

[29] W. Heng and S. Cordelia, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3551–3558.

[30] P. Chakravarty and T. Tuytelaars, "Cross-modal supervision for learning active speaker detection in video," in *Proceedings of ECCV*, 2016, pp. 285–301.

[31] A. Garg, V. Pavlovic, and J. M. Rehg, "Audio-visual speaker detection using dynamic bayesian networks," in *Proceedings of IEEE FG*, 2000, pp. 384–390.

[32] G. Friedland, H. Hung, and C. Yeo, "Multi-modal speaker diarization of real-world meetings using compressed-domain video features," in *Proceedings of ICASSP*, 2009, pp. 4069–4072.

[33] B. G. Gebre, P. Wittenburg, S. Drude, M. Huijbregts, and T. Heskes, "Speaker diarization using gesture and speech," in *Proceedings of Interspeech*, 2014.

[34] E. D'Arca, N. Robertson, and J. Hopgood, "Robust indoor speaker recognition in a network of audio and video sensors," *Signal Processing*, vol. 129, 2016.

[35] P. Chakravarty, J. Zegers, T. Tuytelaars, and H. V. hamme, "Active speaker detection with audio-visual co-training," in *Proceedings of ACM ICMI*, 2016, pp. 312–316.

[36] 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*, vol. abs/1706.00079, 2017.

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

[38] E. Gedik and H. Hung, "Personalised models for speech detection from body movements using transductive parameter transfer," *Personal and Ubiquitous Computing*, vol. 21, no. 4, pp. 723–737, 2017.

[39] S. Thermos and G. Potamianos, "Audio-visual speech activity detection in a two-speaker scenario incorporating depth information from a profile or frontal view," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 579–584.

[40] G. Friedland, H. Hung, and C. Yeo, "Multi-modal speaker diarization of real-world meetings using compressed-domain video features," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4069–4072.

[41] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, "Canal9: A database of political debates for analysis of social interactions," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–4.

[42] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1086–1099, 2017.

[43] X. Alameda-Pineda, J. Sanchez-Riera, J. Wienke, V. Franc, J. Čech, K. Kulkarni, A. Deleforge, and R. Horaud, "Ravel: An annotated corpus for training robots with audiovisual abilities," *Journal on Multimodal User Interfaces*, vol. 7, no. 1-2, pp. 79–91, 2013.

[44] I. D. Gebru, S. Ba, G. Evangelidis, and R. Horaud, "Tracking the active speaker based on a joint audio-visual observation model," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 15–21.

[45] A. Deleforge, R. Horaud, Y. Y. Schechner, and L. Girin, "Co-localization of audio sources in images using binaural features and locally-linear regression," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 718–731, April 2015.

[46] V. Peruffo Minotto, C. Rosito Jung, and B. Lee, "Multimodal multichannel on-line speaker diarization using sensor fusion through svm," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1694–1705, Oct 2015.

[47] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "Multi-speaker tracking from an audio-visual sensing device," *IEEE Transactions on Multimedia*, 2019.

[48] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi *et al.*, "Ava-activespeaker: An audio-visual dataset for active speaker detection," *arXiv preprint arXiv:1901.01342*, 2019.

[49] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: a professional framework for multimodality research," in *Proceedings of LREC*, 2006.

[50] C. Beyan and R. Fisher, "Classifying imbalanced data sets using similarity based hierarchical decomposition," *Pattern Recognition*, vol. 48, no. 5, pp. 1653–1672, 2015.

[51] M. L. McHugh, "Interrater reliability: the kappa statistic," in *Biochemia medica*, 2012.

[52] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proceedings of CVPR*, 2016.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE CVPR*, 2016, pp. 770–778.

[54] M. Shahid, C. Beyan, and V. Murino, "Comparisons of visual activity primitives for voice activity detection," in *Proceedings of 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,*, 2019, pp. 48–59.

[55] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," *Proceedings of IEEE ICCV*, pp. 2960–2967, 2013.

Cigdem Beyan received her Ph.D. degree in Informatics from University of Edinburgh, UK in 2015. She is currently a Postdoctoral Researcher at the Istituto Italiano di Tecnologia, Genoa, Italy in the department of Pattern Analysis and Computer Vision. She has co-authored over 30 papers published in refereed journals and international conferences. Among her main research interest, there are social signal processing, multimodal data analysis, human/animal behavior understanding, deep learning and classification of imbalanced data. She is a reviewer of the most significant multimedia, affective computing, computer vision and pattern recognition journals including multiple IEEE Transactions, and IEEE/ACM conferences. She is a Guest Co-Editor of a special issue in Frontiers in Robotics and AI and in the Editorial Board of ICES Journal of Marine Science covering area of applications of computer vision and machine learning in marine science. She is a member of IEEE since 2013 and an Associate Fellow of the Higher Education Academy UK since 2014.

Muhammad Shahid received his B.S. degree (2010) in Telecommunication Engineering and M.S. degree (2016) in Computer Engineering with major in computer vision from Pakistan. He is currently pursuing his Ph.D. in Istituto Italiano di Tecnologia (IIT), Genoa, Italy in the department of Pattern Analysis and Computer Vision (PAVIS) with the collaboration of Department of Naval, Electrical, Electronic and Telecommunications Engineering (DITEN) from University of Genoa, Italy. His current research includes nonverbal communication, human behavior understanding, social signal processing, multimodal data fusion and deep learning.

Vittorio Murino received the PhD degree in electronic engineering and computer science at the University of Genova, Italy, in 1993. He is currently full professor with the University of Verona, Italy, and a senior video intelligence expert at Huawei Technologies in Dublin, Ireland. In 1995-1998, he was assistant professor with the Dept. of Mathematics and Computer Science of the University of Udine, Italy, and since 1998 he has been working at the University of Verona. He was the chairman of Department of Computer Science from 2001 to 2007, and the coordinator of PhD program in Computer Science from 1999 to 2003. He is a scientific responsible of several national and European projects, and evaluator of EU project proposals. From 2009 to 2019, he worked at the Istituto Italiano di Tecnologia in Genova as the director of Pattern Analysis and Computer Vision Department. His main research interests include techniques for image and video processing for (human) behavior analysis and related applications such as video surveillance, and biomedical imaging. He is the co-author of more than 400 papers published in refereed journals and international conferences, member of the technical committees of important conferences (CVPR, ICCV, ECCV, ICPR, ICIP, etc.), and is guest co-editor of special issues in relevant scientific journals. He is member of the editorial board of Computer Vision and Image Understanding, Machine Vision & Applications, and Pattern Analysis and Applications journals. He is a senior member of the IEEE and a fellow of the IAPR.