Voice Activity Detection by Upper Body Motion Analysis and Unsupervised Domain Adaptation

Muhammad Shahid*, Cigdem Beyan*, and Vittorio Murino
Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, Genova, Italy
{Shahid.Muhammad, Cigdem.Beyan, Vittorio.Murino}@iit.it

Abstract

We present a novel vision-based voice activity detection (VAD) method that relies only on automatic upper body motion (UBM) analysis. Traditionally, VAD is performed using audio features only, but the use of visual cues instead of audio can be desirable especially when audio is not available such as due to technical, ethical or legal issues. Psychology literature confirms that the way people move while speaking is different from while they are not speaking. This motivates us to claim that an effective representation of UBM can be used to detect "Who is Speaking and When". On the other hand, the way people move during their speech varies a lot from culture to culture, and even person to person in the same culture. This results in unrelated UBM representations, such that the distribution of training and test data becomes disparate. To overcome this, we combine stacked sparse autoencoders and simple subspace alignment methods while a classifier is jointly learned using the VAD labels of the training data only. This yields new domain invariant feature representations for training and test data, showing improved VAD results. Our approach is applicable to any person without requiring re-training. The tests applied on a publicly available real-life VAD dataset show better results as compared to the state-of-the-art video-only VAD methods. Moreover, the ablation study justifies the superiority of the proposed method and demonstrates the positive contribution of each component.

1. Introduction

Voice Activity Detection (VAD) is automatically recognizing whether a person is speaking or not in an audio/video recording. VAD is very useful for a number of applications e.g., for the analysis of human-human interaction, human-computer (robot) interaction (HCI), and several industrial applications. For instance, VAD can be used to extract speaker turn features to perform nonverbal social interaction analyses between people [3]. In HCI, having an accurate VAD can allow computer to respond to a specific interlocutor when there is more than one person in the interaction environment [8]. Similarly, video conferencing systems could use VAD to transmit the video of the speaking person only during multi-person meetings. Video navigation and retrieval, speaker model adaptation to enhance speaker recognition, and speaker attributed speech-to-text transcription are other possible applications that VAD can contribute [21].

The most common way to perform VAD is based on audio processing (typically referred to as speaker diarization [42]) while multimodal approaches (generally called active speaker detection) has recently become more popular due to their more accurate performances (e.g., [43, 11]). Multimodal active speaker detection techniques can be based on joint modeling of speech, facial and body cues [11] or can be based on speaker diarization while video is mainly used to track/ localize/ associate the person to the speech [23]. On the other hand, VAD based on visual cues only might be desirable when the audio is not available due to technical or privacy related reasons. Additionally, in social gatherings, such as a cocktail party or a panel, there can be much background noise, which makes the task of distinguishing voices robustly a very challenging problem. However, there are relatively less studies (e.g., [30, 12, 7, 8]) performing VAD with solely visual cues and better performing methodologies are needed.

This paper aims at detecting "Who is Speaking and When" using visual activity-based cues only. We hypothesize that UBM of an active speaker is different from a person who is not speaking and an effective representation of UBM can be used to detect the (non)speaker. This claim is motivated by several findings. First of all, during social interactions individuals unintentionally synchronize their nonverbal and linguistic behavior [35]. Second, gestures occur mainly during speech (with a delay of milliseconds) [17]. Third, it was shown that during narrations about 90% of the time all gestures occurs while person is speaking

*Muhammad Shahid and Cigdem Beyan have equally contributed to this paper.

[36] and it was observed that in a meeting the occurrence of UBM while speaking was for more than 80% of the total speaking time [6]. Fourth, several approaches exploited the relationship between speech and body cues (body motion, hand gestures, mouth movement, fidgeting, body pose, etc.). For instance, in [34], the synchronization between pitch and gestures were used to obtain more human-like artificial agents. Additionally, it has recently been shown that nonverbal behaviours extracted from UBM can result in useful and robust features, e.g., to estimate emergent leaders, and to detect the personality traits in conversations [4]. Such research proves that it is not necessary to extract semantically high-level information from verbal cues.

In this study, we introduce a novel VAD methodology using UBM. Our method extracts features using a Convolutional Neural Network (CNN), trained once, and is applicable for any (new) person without requiring re-training. In other words, instead of estimating hand-crafted features (as the state-of-the-art (SOA) visual-VAD studies performed), this study investigates an end-to-end training, where UBM features are directly learned from the data itself, while the data has been exploited in dynamic images form [5]. As many times mentioned in the psychology literature, the way people move while speaking varies a lot from person to person. This yields dissimilar UBM representations, thus, the distribution of training data and the test data belonging to the new person can be different from each other (a problem called domain-shift). To overcome this, our method presents an unsupervised domain adaptation solution as follows. The new feature representations are obtained by applying two-layers stacked sparse autoencoder, and a simple subspace alignment method, while a classifier is jointly trained using the VAD labels of the training data only. This trained classifier is used to perform VAD for the new test person. By using the proposed domain adaptation technique, which is novel in VAD context, we are not learning person-specific features. We obtain a common representation between the training domain and the new person's domain to perform an effective VAD. Therefore, our method is still a generic approach (independent to person, i.e., not using any VAD labels of the new person).

The contributions of this work are:
a) This is the first attempt that dynamic images [5] are used for VAD task: the full UBM based features are extracted using a CNN model, that is fine-tuned with dynamic images. The resulting nonverbal features are novel and already perform better than SOA features.
b) The VAD results are improved by integrating stacked sparse autoencoder and a simple subspace alignment method to perform unsupervised domain adaptation, which does not require any labels belonging to the test data and supports person-invariant training. This also provides more consistent VAD results such that the detection performances are equally well for all persons.
c) A comprehensive survey on visual data and/or body motion-based VAD are presented.

The rest of this paper is organized as follows. The previous VAD approaches based on visual data and body motion-based cues, including multimodal systems using visual data, are reviewed in Section 2, and the main differences between our work and theirs are highlighted. In Section 3, the details of the proposed methodology are described. The experimental setting is given in Section 4 with a brief description of the dataset used. Following that, in Section 5, we compare the results of our method with the baselines, and we perform an ablation study to show the importance of each component of our method. Finally, we conclude the paper with a summary and list the future work in Section 6.

2. Related Work

As an earlier work on computer vision-based active speaker detection, Rehg et al. [37] proposed a Bayes Net model, which combines face detection, skin color, skin texture and mouth motion sensors. VAD using features extracted from face (e.g.; face gestures, lip movements, head movements) is still an active area. For instance, in [31] facial movements have been detected using Spatiotemporal Gabor filters applied to the mouth, the head or the entire face, while mouth region gave the best result. Later, in [27], head and lip movements have been used to detect and localize the active speaker in a human-machine multiparty interactive dialogue setting. In detail, the head movement versus fusion of head and lip movements and lip movement versus fusion of lip and head movements were analyzed in three settings: speaker dependent, speaker independent and hybrid. It was observed that head movements contributes significantly towards VAD and outperforms lip movements except speaker independent settings, and fusion of both improves performance of VAD [27]. More recently, face features extracted from AlexNet to perform VAD in real-time multiparty interactions, was presented in [39]. Long short-term memory (LSTM) has been used to model the temporal dependencies between face features over time with the VAD labels extracted from an acoustic speaker diarization method. Then, the trained LSTM model has been used to predict if a given frame composed of face, is speaking or not [39].

There are also methods using visual activity (e.g.; head activity, hand gestures, full body or UBM) for VAD. In [30], the correlation between head/ hand activity and speaking status was analyzed by assuming that; the speaker is the one who moves most, and group's visual attention is more likely to converge on the speaker than on others. In that study [30], visual activity of skin-colour regions has been represented using Discrete Cosine Transform (DCT) coefficients and residual coding bit-rate, while a Bayesian approach has

been used to detect visual focus of attention (VFOA). Supervised and unsupervised methods have been applied to test the features extracted for VAD in meetings [30]. Even though, that study [30] showed promising results, detection of VFOA is mostly robust when there are multiple cameras capturing each person individually at close distance. Gebre et al. [22] used motion history images (MHI) as a likelihood measure of speaking activity, and their method showed encouraging results on the same meeting dataset used in [30]. Cristani et al. [12] utilized the relationship between speech and gestures to detect the active speaker in surveillance scenarios. That method [12] is based on a local video descriptor, which extracts the optical flow of human body, and encodes optical flow energy and complexity using an entropy-like measure [12]. The results of that study [12] are promising, but the dataset they used has a top-view, which already decreases the possibility of occlusions and also the frames where the region of interests overlap (i.e., inter-person occlusions) were discarded from the analyses. Latterly, directional audio information has been used to label improved trajectory features extracted from head and torso tracks of people as speaking/non-speaking [7]. These labels have been used for the training of an SVM to perform video-based VAD. Improved trajectory has been calculated from 15 consecutive frames, pooled by a Fisher vector (FV) representation and has been associated with spatio-temporal features e.g., the mean pixel location of the trajectory, and Histogram of Gradients (HoG), Histogram of Flow (HoF) and Motion Boundary Histogram (MBH) features. Chakravarty et al. [8] extended that scheme [7] to an online learning setting, starting from a generic active speaker detection classifier, which gradually adapts itself to a specific person.

One of the first audio-visual VAD approach was presented in [19], which was tested on human-centered user-interfaces. In that study [19], face, skin, texture, mouth motion, and silence detectors have been optimally fused with contextual information using a Dynamic Bayesian Network (DBN) architecture, showing improved performance as exploiting the temporal correlation between audio and visual sensors. Graphical models have been also used e.g., in [1, 10], such that the audio from single microphone source have been used to determine if someone is speaking, and visual features have been used to localize the active speaker. In [23], the results of multiple audio-based source localization techniques have been combined with a tracker allowing to associate multiple persons to the multiple speeches. Similarly, in [29], faces have been mapped with voices based on the correlation between speech and face clusters. That method [29] is advantageous as not requiring any a priori but it requires an accurate pre-trained face and speech detector. In [15], tightly cropped face images and sound mixtures have been jointly modeled with a CNN and a bidirectional LSTM model to perform speaker-independent VAD. In [13], active speakers were found by tracking and recognizing voice of people with a hierarchical audio-visual system applied to surveillance scenarios. In that study [13], by using multiple modalities, challenges; large occlusions and cross-talks were handled. However, that approach is limited since only the most dominant speaker could be detected and tracked but the other persons speaking at the same time with the dominant speaker, cannot be detected.

The first approach using entire body motion information together with audio features was presented in [43], which was tested on meetings having single stationary camera and a single microphone. That approach [43] is based on long term co-occurrences between audio and video subspaces found by clustering, does not require training, and does not rely on a priori. However, it is not clear if that method [43] is able to detect overlapping active speakers. The acoustic features; MFCCs and visual activity-based features; the average motion vector magnitude in skin blocks have been fused in [18]. These multimodal features showed improved results as compared to audio-only baseline [18]. In [20], speaker models have been learned from speech samples corresponding to gestures such that the occurrence of gestures indicates the presence of speech and the location of gestures indicates the identity of the speaker.

In [9], cross-modal supervision from video within an audio-visual co-training has been addressed. In detail, a generic body cues-based VAD classifier trained by directional audio, has been used to train a video-based person-specific VAD. Then, learned video classifiers have been used to supervise the training of personalized voice models. The drawback of that study [9] is, performing person-specific VAD, which requires training data for each new person VAD is performed. Recently, in [40], a CNN model has been used to learn features from facial area while acoustic data has been represented a with Mel-filterbank features. The features coming from two-modalities have been concatenated and used by unidirectional LSTM. In that study [40], joint modeling was not performed, which is different from [11] that proposed a two-stream CNN model that learns an embedding between the sound and the mouth motions. The results in [11] showed that joint learning of audio and synchronized lip-motion could improve the active speaker detection results as compared to visual activity-based VAD presented in [8].

Unlike any work discussed here, wearable sensors (a single triaxial accelerometer worn around the neck) have been used to perform VAD in crowded scenes in [24]. In that work [24], body movements have been represented in terms of power spectral density of a motion signal and transductive transfer learning has been applied to be able to better model individual differences in speaking behavior of different persons. As the first attempt for speech activity de-

tection, depth visual information has been combined with audio and planar video information in [41]. The results showed that depth information significantly contributes to VAD. However, that model [41] was tested on simple scenarios having two persons, and should be tested on more realistic interactions to validate the results.

2.1. Highlights of The Proposed Method

Unlike multimodal VAD techniques discussed above, we only utilize visual cues. We neither analyze the head motions like in [31, 27, 30] nor the face features as in [31, 27, 39]. Detecting visual focus of attention (VFOA) as applied in [30] is also out of the focus of this study, since it is not always true that the majority of the persons are gazing the speaker (for instance, in panel discussions when panelists are sitting in a single row, it is rare that they face each other). Dissimilar to our approach, there are a lot of methods based on lip movements e.g.; [31, 27, 11, 33]. However, these techniques are limited as detecting lip motion is not always possible. For instance, 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 the speaker's face is occluded by her hands, facial features detectors fail to detect the lips. Additionally, in [7], it is also shown that body activity-based features can outperform lip motion-based features for VAD. We only use the features representing upper body activity of a person, which is similar to [12, 22, 7, 8]. However, the way we initially represent the body motion (by dynamic images [5]), and then model and extract features (by using an end-to-end deep learning approach) are completely different from these works.

Our method is a generic approach such that it does not require any VAD labels belonging to the person in the test data. Also, it does not need any video frame of the test data for the training of the feature extractor. This is advantageous as compared to the studies, e.g.; [8, 9], and the majority of the multimodal approaches, which presented a person-specific VAD. The way people gesticulate while speaking varies a lot from person to person, thus, a person-specific model can outperform a generic model [8]. However, person-specific models are restrictive as they need to be re-trained for each new person. Moreover, person-specific models can still have data discrepancy problem, which results in poor VAD performance. To handle data discrepancy problem, unlike applying online training as performed in [8], we present an unsupervised domain adaptation method, which have never been applied for video-based VAD. That domain adaptation method not only improves the VAD performance on average, but also provides consistent results such that VAD performances are equally well for all persons.

The dataset we use to evaluate our method is from a real-life panel i.e, not a role-play scenario based small group meeting as in [30, 18, 20, 22, 43]. In such meeting datasets, all participants know what the group task is, the cameras are always static, there are more than one cameras capturing participants from their frontal view, and the places of the cameras are known by the participants. These results in less challenging head, body and even lip motion detection. [8] is one of the baseline work as utilizing the same dataset with us. One of the main difference between our study and theirs is that, all the features they proposed were hand-crafted, while ours are extracted from the data itself, which are based on deep learning. In that study [8], temporal continuity (is based on the heuristics that if a person is speaking it is more likely that she will continue speaking for a while rather than stop speaking), was used and the misclassification results were largely corrected. However, it is not clear how the window size of the temporal continuity should be selected to obtain accurate VAD results. Instead, our method does not need to apply temporal continuity to correct the results. The results of [8] are also not stable, such that VAD results are good for some persons, while for others they are not sufficient.

3. Proposed Method

The proposed method is illustrated in Figure 1. First, multiple dynamic images, representing UBM of the person in the given video, are generated from the training and test videos individually (Section 3.1). Then, ResNet50 [28] is fine-tuned for VAD task while the dynamic images of training data are the inputs. This fine-tuned model is used to extract features for each dynamic image (Section 3.2) of training and test data. Following that, the features obtained for training data could be used to train a classifier (such as Support Vector Machine as applied in this study) or even without extracting features, the trained model could be used with the softmax function to classify the dynamic images of the test data as speaking or not-speaking (i.e., a fully end-to-end system). However, the softmax result (referred as $Softmax$ in Section 5) and the classifier's result when features extracted were the input (referred as SVM in Section 5) showed that, while these approaches work well for some speakers, their performance is not sufficient for some other speakers. This might be due to the fact that the way people move while speaking varies from person to person [7, 8], which results in dissimilar UBM representations (in our case dissimilar dynamic images) that might cause a domain shift problem resulting in lower classification performance. This challenge is overcome by applying an unsupervised domain adaptation method, and a SVM classifier.

The parameters of the domain adaptation method and the SVM are jointly learned. The proposed domain adaptation method aligns the subspace of train (source domain) and test (target domain) features extracted from fine-tuned ResNet50 model [28] after applying two-layers stacked

sparse autoencoder (Section 3.3). An SVM classifier is trained using the new feature representations of training data obtained from the domain adaptation approach, and the corresponding VAD labels. To determine the voice activity of a test dynamic image, the trained SVM classifier (Section 3.4) is applied to the corresponding data represented in terms of the new domain adapted features, resulting in a VAD label as speaking or not-speaking. This predicted label corresponds to the test video frames, those the test dynamic image is constructed from.

3.1. Multiple Dynamic Images Construction

There are diverse way to detect the visual activity (VisualAct) of a person, which can be used to extract UBM. For example, in [30], a combination of motion vectors, DCT coefficients and residual coding bit-rate were used to estimate VisualAct for VAD. In [22], motion history images (MHI) were used to represent the VisualAct for video-based VAD. Optical flow is another popular method used to extract VisualAct to detect the speakers [12]. Recently, to summarize the short-term spatio-temporal content of a video in a single image, Bilen et al. [5] proposed dynamic image representation, which achieved significantly better results for activity recognition. Dynamic image can be seen as a compact representation of a video segment, which summarizes the appearance and motions of it. Construction of a dynamic image 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 [5].

We obtain multiple dynamic images to represent an input video. For each consecutive 10 frames in a given video, we obtain one dynamic image without overlapping. The number of frames used, are defined arbitrarily. We observed that dynamic images constructed from RGB domain (raw video frames) were good at capturing the motions belong to the person in the video. Some example dynamic images from the dataset used, are given in Figure 2.

3.2. ResNet50 Fine-tuning and Feature Extraction

Dynamic images can be used with any CNN architectures for fine-tuning as also shown in [5]. We first fine-tuned AlexNet (pre-trained on ImageNet dataset). However, this resulted in over-fitting 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 might have better feature representation capacity [26], we also performed our analysis by using ResNet50 [28], which gave significantly better (p-value < 0.01) VAD results as compared to AlexNet.

When dynamic images extracted from speaking and non-speaking video segments are the inputs, ResNet50 (pre-trained on ImageNet dataset) is fine-tuned by adding a fully-connected layer after the final convolution layer (called $fc1$ in the rest of this paper). This $fc1$ layer has 2048 neurons and its weights are randomly initialized. Only the weights of fully connected layer are updated during fine-tuning, in other words the weights of convolution layers are not updated. The model is trained in end-to-end manner with cross entropy loss function, Adam optimizer, $10e^{-5}$ learning rate and for 20 epochs.

During fine-tuning, we noticed that there is much more non-speaking segments as compared to the speaking segments, but training with an imbalanced data misleads the classification task [2]. To overcome this, the data in each batch (in total 128 samples) is balanced such that 64 speaking and 64 non-speaking randomly selected samples are used. Additionally, data augmentation is applied as follows. Some randomly selected training images are horizontally flipped and/or a 64x64 randomly selected patch is replaced with the mean value of the images, which can be seen as a dropout in input layer.

3.3. Unsupervised Domain Adaptation

The autoencoder was introduced by Rumelhart et al. [38] as an unsupervised learning model and have been used for different purposes. It is essentially a neural network trained to map the input to an output, which is a reconstruction of the input. The simplest form of an autoencoder has a single hidden layer that encodes an input x to its new representation y, with an activation function f (usually non-linear), weight matrix (W) and a bias vector b. The encoded y is then decoded to reconstruct the input x (shown as x_r below). Training is based on a loss function that is minimized while hidden/output layer weight matrices (W and W') and hidden/output layers bias vectors (b and b') are optimized. These can be summarized as follows.

$$y = f(Wx + b) \tag{1}$$

$$x_r = f(W'y + b') \tag{2}$$

Recently, the autoencoder is used for unsupervised domain adaptation [32] such that a common representation between training (source domain) and test (target domain) data is learned. This idea has never been applied for video-based VAD. Previous studies [25, 14] showed that, it is possible to learn better performing domain-invariant features thanks to non-linear transformation property of autoencoder. However, using autoencoder only might not be sufficient (i.e.; does not always grantee effective features), and in some cases autoencoder can result in more discrepancy across target and source domain. We combine autoencoder with a subspace alignment approach (referred as $AE + SA$ for the rest of the paper). The autoencoder structure and the subspace alignment technique used are summarized as follows.

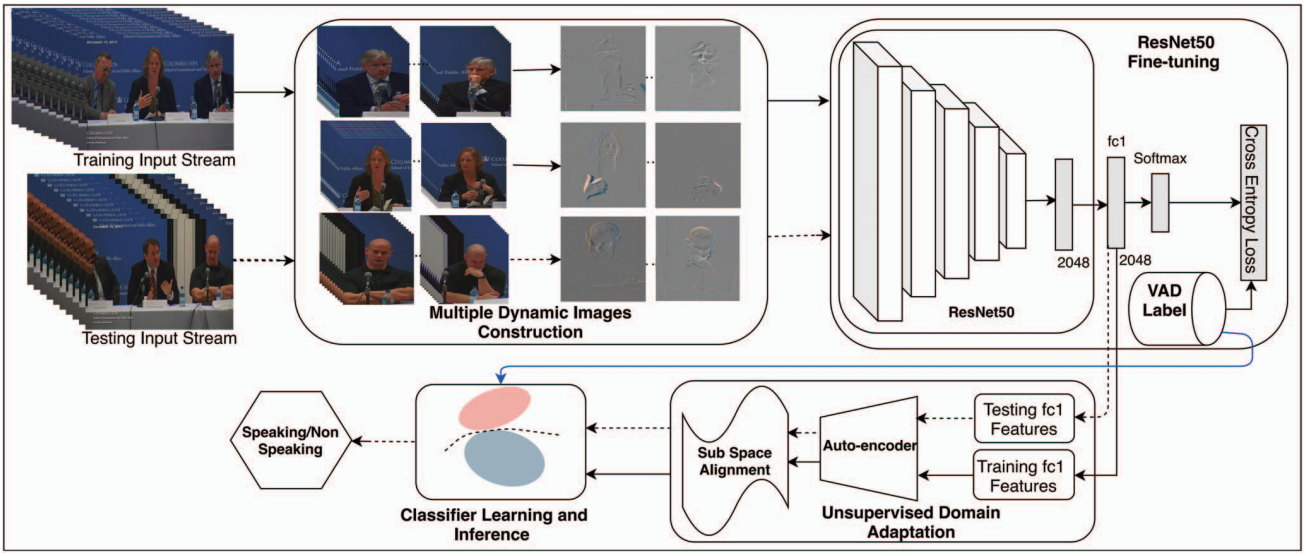

Figure 1. The illustration of the proposed method.

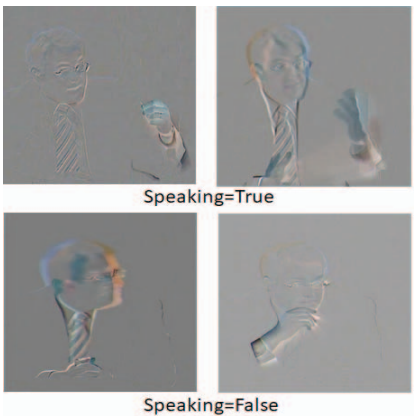

Figure 2. Example dynamic images obtained from 10 consecutive speaking and not-speaking frames.

Stacked Sparse Autoencoder: A two-layers stacked sparse autoencoder ($2AE$) is used. The number of hidden neurons in each layer, $l2$ weight normalization factor (i.e., regularization parameter for weight updating), sparsity constraints for sparsity control, and sparsity proportion are all optimized layer-wise. In detail, the number of neurons in first and second layers are set to 512 and 128, respectively while the $l2$ weight regularization is 0.002, and sparsity regularization is four. The sparsity proportion in each encoding layers are 0.15 in first layer and 0.40 in second layer. These values correspond to the best VAD performance obtained for the validation set. The training of $2AE$ is performed for 300 epochs using complete test data and the same amount of randomly selected training samples (called balanced training in Section 5). The loss function contains two parts; a mean square error between input and output, and two regularization terms. The first regularization term (weight decay) controls the overfitting and the second regularization is sparsity constraint on the hidden units.

Principal component analysis based subspace alignment (PCA-SA) [16]: Suppose, S_f is the feature set of source data and T_f is the feature set of target data, both having d dimensions. First, the source and target features are normalized to zero mean and unit variance. Then, PCA is applied to both data with N eigenvectors corresponding to top N eigenvalues. These N eigenvectors are considered as the bases of source and target subspaces represented as U_{f_S} and U_{f_T} (where $U_{f_S}, U_{f_T} \in R^{dxN}$). As U_{f_S} and U_{f_T} are extracted through singular value decomposition (SVD), they are orthonormal to their transposed form ($U'_{f_S}U_{f_S} = I, U'_{f_T}U_{f_T} = I$). Later, a linear transformation matrix M, which transforms the source subspace coordinates U_{f_S} into target feature subspace U_{f_T}, is learned by optimizing the Bregman divergence as follows.

$$F(M) = \|U_{f_S}M - U_{f_T}\|_F^2 \tag{3}$$

$$M^* = \text{argmin}_M(F(M)) \tag{4}$$

where $\|.\|_F^2$ is the Frobenius norm. Eq. 3 can be re-written as:

$$F(M) = \left\|U'_{f_S}U_{f_S}M - U'_{f_S}U_{f_T}\right\|_F^2 = \left\|M - U'_{f_S}U_{f_T}\right\|_F^2 \tag{5}$$

which results in $M^* = U'_{f_S}U_{f_T}$ and $U_c = U_{f_S}M^*$, where U_c stands for target-source common coordinate system (called the target aligned source coordinate system in [16]). As a result of this alignment, the new feature representations of target (f_{t_T}) and source (f_{t_S}) domains are found as $f_{t_T} = T_f U_{f_T}$ and $f_{t_S} = S_f U_c$, respectively. The

only parameter is the number of eigenvectors, whose value is found automatically from the set of values: $\{5, 10, 15, 20, 25, 30\}$, based on the best VAD performance obtained for the validation set.

3.4. Classifier Learning and Inference

The proposed method could be combined with any classifier for learning and inference. We use linear Support Vector Machine (SVM), which is in line with many studies such as [7, 8]. As kernel parameter C was taken as 10^k while $k = \{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$. The SVM model is jointly trained with $2AE + SA$ and is used for the classification of test data.

4. Experiments

The proposed method was evaluated using the only publicly available real-life VAD dataset, called Columbia [8]. This dataset contains a 87 minutes-long video (frame rate is 30 frames per second), which is from a panel discussion at Columbia University. There are seven participants on the panel. The field of the view of the camera is changing to focus on smaller groups of panelist at a time. Following the SOA [8, 9, 11], 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 participants (Bell, Bollinger, Lieberman, Long, Sick) while two-three participants are visible at any one time. We used the VAD labels, speaking/not-speaking for each video frame, belonging to these five persons, to be able to compare our results with SOA. This dataset supplies the bounding boxes i.e., the detections of each person. However, we observed that these bounding boxes contain only the head of the panelists, instead of the whole upper body. Therefore, we re-extracted the bounding boxes, this time containing the whole upper body of a given panelist, which will be supplied upon request. Leave-one-person-out cross validation and F1-score as the evaluation metric were used to compare our results with SOA [8, 9, 11].

5. Results

The best results of the SOA [8, 9, 11] and the proposed method are given in Table 1. The results of the ablation study is also reported in the same table. Ablation study allows us to make a performance comparison particularly between: *i)* AlexNet and ResNet50, *ii)* ResNet50 w/ and w/o data augmentation, *iii)* one-layer sparse autoencoder and two-layers stacked sparse autoencoder, *iv)* sparse autoencoder w/ and w/o balanced training, *v)* the proposed method and the proposed method without sparse autoencoder i.e. using subspace alignment only for unsupervised domain adaptation, *vi)* the proposed method and the proposed method without subspace alignment i.e. using sparse autoencoder only for unsupervised domain adaptation.

As seen (Table 1), the average performance of the proposed method is better than the video-only SOA VAD approach [8]. The performance of [8] is highly dependent to the choice of window size (W) of temporal continuity algorithm (see Section 2.1 for more information). Given that we create dynamic images for each 10 consecutive frames, it can be more fair to compare the performance of the proposed method with [8] while W is equal to 10. In this case, the proposed method performs even better (11%) than [8]. The same arguments are correct for [9], when its video-only features are used. Additionally, the performance of the proposed method is as good as the SOA multimodal VAD approach [11]. This is an important achievement since the proposed method is based on visual activity only, while [11] utilizes lip motions with audio.

Better average VAD performance of the proposed method is definitely very important but having low VAD standard deviation (STD) of all participants (while still performing better on average), is also a significant aspect of the proposed method. In detail, the performance of [8] has fluctuations such that it performs well for some persons (e.g., Long: 86.90%), while performs highly worse for some others (e.g., Bollinger: 65.89%). This can be observed from the high STD values: 8.45% and 10.36% as well. The same arguments are correct for [9], when its video-only features are used. On the other hand, [11] has much lower STD (5.94%) than [8, 9], which is as good as the proposed method. The performance of the proposed method is the most consistent (STD= 5.14%). This is not only due to applying unsupervised domain adaptation, but also due to the superiority of the proposed features ($fc1$ features of ResNet50) as compared to the SOA features. This can be seen from STD values: 7.82% (Softmax), 7.76% (Softmax), 6.97% (SVM), i.e.; all of them are much lower than STDs of [8] while all of them also perform better than [8], on average. Once the autoencoder and subspace alignment techniques are integrated individually, the average performance gradually improves (88.31% and 87.43%, respectively), while STD values decreases (5.18% and 6.56%, respectively).

These results also proves that, the way ResNet50 is fine-tuned clearly outperforms fine-tuning AlexNet. Data augmentation (see Section 3.2) applied during the fine-tuning of ResNet50, improves the VAD performance for all participants. Balanced training of sparse autoencoder (see Section 3.3) contributes positively to the VAD performances for all participants independent to the number of layers (one-layer or two-layers stacked sparse autoencoder). Moreover, two-layers stacked autoencoder enhances the VAD performance, which also results in lower-dimensional feature space as compared to single layer autoencoder. The

Table 1. F1-scores (%) on the Columbia dataset. The results of [8] is taken from [11]. AVG and STD stand for average and standard deviation of F1-scores of all participants, respectively. W, FT, w/AUG, AE, $2AE$, BT and SA stand for window size, fine-tuning, with augmentation, sparse autoencoder, stacked sparse autoencoder, balanced training and subspace alignment, respectively. The best results of all are emphasized in bold-face.

Method	Bell	Bollinger	Lieberman	Long	Sick	AVG	STD	Details
[8, 9]	82.90	65.80	73.60	86.90	81.80	78.20	8.45	W=10, video-only results of [9]
[8, 9]	90.30	69.00	82.40	96.00	**89.30**	85.40	10.36	W=100, video-only results of [9]
[11]	**93.70**	83.40	86.80	**97.70**	86.10	**89.54**	5.94	W=10
Softmax	78.29	84.38	59.39	63.59	64.14	69.96	10.76	FT AlexNet
Softmax	85.95	91.08	90.71	71.84	85.95	85.11	7.82	FT ResNet50
Softmax	86.07	93.30	91.88	73.62	86.34	86.24	7.76	FT ResNet50 w/AUG
SVM	86.35	93.78	**92.34**	76.09	86.25	86.96	6.97	FT ResNet50 w/AUG, $fc1$ fea.
AE+SVM	86.54	92.95	92.19	77.57	86.96	87.24	6.15	FT ResNet50 w/AUG, $fc1$ fea., 1-layer AE.
AE+SVM	87.18	93.58	92.10	78.22	87.39	87.69	6.00	FT ResNet50 w/AUG, $fc1$ fea., BT 1-layer AE.
2AE+SVM	86.34	94.44	92.09	78.66	87.16	87.74	6.09	FT ResNet50 w/AUG, $fc1$ fea., 2-layers AE.
2AE+SVM	87.28	94.01	92.20	80.70	87.35	88.31	5.18	FT ResNet50 w/AUG, $fc1$ fea., BT 2-layers AE.
SA+SVM	86.51	94.12	**92.33**	77.32	86.87	87.43	6.56	FT ResNet50 w/AUG, $fc1$ fea., $PCA-SA$.
Proposed Method (2AE+SA+SVM)	87.28	**96.35**	92.15	83.03	87.21	**89.20**	**5.14**	FT ResNet50 w/AUG, $fc1$ fea., BT 2-layers AE, $PCA-SA$

proposed method performs well when $PCA-SA$ is used, which is a very simple method requiring few parameters to be learned. $PCA-SA$ is also useful to show the goodness of the proposed method and the proposed features without the necessity to resort to more complex algorithms for UBM-based VAD. When domain adaptation is applied ($2AE+SA+SVM$), for "LONG" significantly better result (p-value < 0.01) as compared to SVM was obtained.

6. Conclusions

We have demonstrated that computer vision and deep learning-based upper body motion (UBM) analysis is effective for voice activity detection (VAD) task, which can be especially important in case other modalities such as audio is neither feasible to acquire nor reliable. We have utilized dynamic image representation with an end-to-end deep learning-based methodology to extract novel features from whole upper body. This is different from many previous works that focused on motion in individual body parts like lips, face or head, and also unlike the works being based on feature engineering. Additionally, we have interoperated the challenge of domain specificity, which leads to performance degradation across participants depending upon the characteristics of each person (i.e. some person moves more than others) by proposing an effective domain adaptation technique.

This is the first attempt that to realize VAD, dynamic images are used for motion representation, and autoencoder is associated with subspace alignment techniques. The proposed method is a generic, person-independent approach, which does not require any VAD labels belong to the person in the test. This has been rarely realized in previous approaches.

The comparisons between the proposed method and the SOA video-only VAD methods showed better performance of the proposed method while it performed as well as SOA multimodal VAD approach when an unconstrained real-life panel discussion VAD dataset is used. Moreover, unsupervised domain adaptation provided more consistent VAD performance such that the detector works equally well for all participants and for some participants significantly improved results were obtained as compared to not applying domain-adaptation. The domain-adaptation part of our method will be better investigated, once a dataset containing participants having different ethnic origins, is collected. The proposed method can be applied to any dataset since it is not based on specific body part detection e.g. lips, face. In case background motion exists, person detection can be performed before constructing dynamic images for better performance.

A limitation of the proposed method is requiring the number of raw video frames to construct a dynamic image. All other values of parameters are found automatically during training based on the best VAD performance of the validation data. As future work, an automatic way to construct dynamic images in necessary number will be explored. Additionally, the proposed method will be adapted to perform VAD in crowd, and multiparty egocentric video streams.

References

[1] M. J. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):828–836, 2003.

[2] C. Beyan and R. Fisher. Classifying imbalanced data sets using similarity based hierarchical decomposition. *Pattern Recognition*, 48(5):1653–1672, 2015.

[3] 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.

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

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

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

[7] 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*, pages 87–90, 2015.

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

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

[10] T. Choudhury, J. M. Rehg, V. Pavlovic, and A. Pentland. Boosting and structure learning in dynamic bayesian networks for audio-visual speaker detection. In *Object recognition supported by user interaction for service robots*, pages 789–794, 2002.

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

[12] 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*, pages 72–80, 2011.

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

[14] J. Deng, Z. Zhang, F. Eyben, and B. Schuller. Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, 21(9):1068–1072, 2014.

[15] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *CoRR*, abs/1804.03619, 2018.

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

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

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

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

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

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

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

[23] 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*, 40(5):1086–1099, 2018.

[24] E. Gedik and H. Hung. Personalised models for speech detection from body movements using transductive parameter transfer. *Personal and Ubiquitous Computing*, 21(4):723–737, 2017.

[25] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of ICML*, pages 513–520, 2011.

[26] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[27] 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)*, pages 1207–1211, 2016.

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

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

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

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

[32] M. Kan, S. Shan, and X. Chen. Bi-shifting auto-encoder for unsupervised domain adaptation. In *Proceedings of ICCV*, pages 3846–3854, 2015.

[33] E. E. Khoury, C. Senac, and P. Joly. Audiovisual diarization of people in video content. *Multimedia Tools and Applications*, 68(3):747–775, 2014.

[34] S. Kopp and I. Wachsmuth. Synthesizing multimodal utterances for conversational agent. *Computer Animation and Virtual Worlds*, 15(1):39–52, 2004.

[35] N. Latif, A. V. Barbosa, E. Vatiokiotis-Bateson, M. S. Castelhano, and K. G. Munhall. Movement coordination during conversation. *PLOS ONE*, 9(8):1–10, 2014.

[36] D. McNeill. So you think gestures are nonverbal. *Psychological review*, 92(3):350–350, 1985.

[37] J. M. Rehg, K. P. Murphy, and P. W. Fieguth. Vision-based speaker detection using bayesian networks. In *Proceedings of IEEE CVPR*, volume 2, pages 110–116, 1999.

[38] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. 323:533–536, 1986.

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

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

[41] 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)*, pages 579–584, 2016.

[42] S. E. Tranter and D. A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557–1565, 2006.

[43] 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*, 18(11):1608–1617, 2008.