Audio-Visual Voice Activity Detection Using Diffusion Maps

David Dov, Ronen Talmon, *Member, IEEE*, and Israel Cohen, *Fellow, IEEE*

***Abstract*—The performance of traditional voice activity detectors significantly deteriorates in the presence of highly nonstationary noise and transient interferences. One solution is to incorporate a video signal which is invariant to the acoustic environment. Although several voice activity detectors based on the video signal were recently presented, merely few detectors which are based on both the audio and the video signals exist in the literature to date. In this paper, we present an audio-visual voice activity detector and show that the incorporation of both audio and video signals is highly beneficial for voice activity detection. The algorithm is based on a supervised learning procedure, and a labeled training data set is considered. The algorithm comprises a feature extraction procedure, where the features are designed to separate speech from nonspeech frames. Diffusion maps is applied separately and similarly to the features of each modality and builds a low dimensional representation. Using the new representation, we propose a measure for voice activity which is based on a supervised learning procedure and the variability between adjacent frames in time. The measures of the two modalities are merged to provide voice activity detection based on both the audio and the video signals. Experimental results demonstrate the improved performance of the proposed algorithm compared to state-of-the-art detectors.**

***Index Terms*—Audio-visual speech processing, diffusion maps, voice activity detection.**

I. Introduction

VOICE activity detection is an essential component in many applications such as speech and speaker recognition [1], speech coding, speech enhancement and dominant speaker identification [2]. Often, voice activity detection algorithms such as those presented in [3], [4], [5], [6], [7] and [8] assume that noise is slowly varying with respect to speech. In the presence of highly non-stationary noise and transients, such as keyboard typing, door knocking, and office noise [9], [10], [11], [12], [13] this assumption does not hold, and the performance of these algorithms significantly deteriorates. A different type of prior assumptions, used for voice activity detection, relate to specific characteristics of the speech signal, e.g. the periodicity of the signal in the frequency domain for voiced phonemes [14] and the tendency to increase one's vocal intensity in the presence of noise with low Signal to Noise Ratio (SNR) [15]. Recently, a Voice Activity Detector (VAD) for non-stationary environments was presented in [16], where transient effects are reduced by averaging estimates of noise statistics over short time windows. This detector is based on a spectral clustering method for the detection of voice activity and was shown to outperform competing state of the art detectors. Yet, its performance in presence of transients is limited, because transients are not estimated along with the noise due to their fast varying nature.

Nowadays, video calls are becoming a standard way to communicate, and modern products, e.g. smartphones and laptops, have integral microphones and cameras. The availability of a video signal, in addition to the audio signal, can be highly beneficial for voice activity detection, especially in challenging acoustic environments, since the video signal is invariant to acoustic noise in general, and transients in particular.

Existing voice activity detection methods, which are based on visual data, focus on the analysis of the region of the mouth, and in particular the lips. However, their main drawback is their dependency on the detection of the lips, which often rely on artificial markers and whose accuracy may be degraded due to skin color or illumination conditions. For example, in the studies presented in [17] and [18], the detection is based on features, which are constructed from contours of the lips. However, the extraction of the contours requires that the lips would be marked using a blue makeup. In [19], the presented VAD exploits the shape and color of the lips, which are obtained assuming that the lips are marked by key points. Another approach for extracting the lips, which assumes that the color of lips is significantly different from the color of skin, was proposed in [20] and [21].

Another approach to visual voice activity detection relies on the dynamics of the region of the mouth. In [19], a second algorithm based on the movements of lips was presented as well. This algorithm focuses on the analysis of the region of the mouth, which is enhanced using a retinal filter. Although exhibiting good performance, the detection was found to be sensitive to lips movements in speech absent intervals. In [22], a similar approach based on motion estimation in the region of mouth was presented. Motion fields are used as features to characterize the change of the position of the mouth over time and a Hidden Markov Model (HMM) is used for the classification. Such a motion estimation approach was also utilized in [23]. There, the energy in the mouth region is defined using optical flow and serves as a feature for a classifier based on an HMM.

Manuscript received July 24, 2014; revised November 26, 2014; accepted February 02, 2015. Date of publication February 19, 2015; date of current version March 06, 2015. This work was supported by the Israel Science Foundation under Grant 1130/11. The work of R. Talmon was supported in part by the European Union's Seventh Framework Programme (FP7) under Marie Curie Grant 630657 and in part by the Horev Fellowship. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Woon-Seng Gan.

The authors are with the Department of Electrical Engineering, The Technion-Israel Institute of Technology, Haifa 32000, Israel (e-mail: davidd@tx.technion.ac.il; ronen@ee.technion.ac.il; icohen@ee.technion.ac.il).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2015.2405481

A VAD based on intensity values in the region of the mouth was presented in [24], where the detection is based on the number of low intensity pixels, which are modeled using a Gaussian model for speech and non-speech hypotheses. Although high detection rates were reported, the algorithm may be limited in real time applications because the entire speech sequence is required in advance to estimate the noise statistics.

Although VADs based on video signals have an advantage over VADs based on audio signals in noisy conditions, they usually fail to compete with audio based detectors, since a trivial classifier based on the energy of the audio signal can obtain near perfect performance in quiet environments. Therefore, a bimodal VAD may combine the advantages of both audio and video signals. A study which compares several Audio-Visual Voice Activity Detectors (AV-VAD) was presented in [25], where the detectors are categorized according to classification and fusion schemes. The AV-VADs are incorporated in a Speech Recognition System (SRS) and are evaluated according to the recognition scores. The highest recognition score was achieved using an AV-VAD where the features are the log of the power of the audio signal and the vertical variance of the optical flow vectors for the video signal. The modalities are fused in the features level using a weighted sum and the combined audio-visual feature is compared to a threshold for the classification. Another approach for AV-VAD which is also designed for incorporation in an SRS was presented in [26]. The audio signal is represented by a feature based on a likelihood score for silence which is evaluated in the SRS based on recognition scores, and the video features are based on the width and the height of the lips. The modalities are merged in the features level, and the classification is based on a supervised learning procedure, which assumes a Gaussian Mixture Model (GMM) in the features domain. Another AV-VAD which is also based on a supervised learning procedure was presented in [27], where the video signal is analyzed using a Bayesian approach to detect the lips, followed by an HMM to model the lips movements. The audio signal, which is assumed to be acquired in a microphone array, is used to compute a spatio-temporal coherence of the source. Then, another HMM is used for speech presence estimation. Finally, the two modalities are combined at the classification stage using a tree based classifier.

In this paper, we present an algorithm for audio-visual voice activity detection. The inputs to the algorithm are audio and video signals recorded in a single microphone and a single video camera, respectively. The algorithm is based on a supervised learning procedure, and we consider a training data set which comprises speech signals contaminated by different types of noise and transients, and is labeled according to the presence and the absence of speech. The algorithm comprises two steps: first, a low dimensional representation of the signals of each modality is constructed by applying dimensionality reduction to high dimensional features of each modality. The audio features are based on weighted Mel-Frequency Cepstral Coefficients (MFCC) [28] and are designed to separate the stationary from the non-stationary parts of the signal. The video features are based on motion vectors, which capture well both the shape of the mouth and its dynamics. By adopting similar concepts to the spectral clustering algorithm presented in [16], we exploit diffusion maps [29], a manifold learning method, which is applied separately and similarly to the features computed from each modality. Diffusion maps provides a low dimensional representation of the signals which is suitable for merging data captured from different types of sensors [30]. In addition, it captures the intrinsic structure of the data and provides a good distance metric to separate speech and non-speech frames. Second, a measure for voice activity is defined based on the diffusion mapping. This measure incorporates both a supervised clustering procedure, which is based on a GMM, and an unsupervised procedure that exploits the variability of consecutive frames. The GMM is used to separate speech and non-speech clusters according to the labeled training data, and the unsupervised procedure separates the two clusters by assuming high variability between adjacent speech frames. The computed measures for voice activity from the two modalities are merged into a single bimodal measure, which is in turn used to estimate the speech presence indicator.

The proposed algorithm is tested in the presence of highly non-stationary noise and transients. Experimental results demonstrate the improved performance of the single modal versions of the proposed algorithm over state-of-the-art single modal VADs. In addition, we show that the proposed AV-VAD outperforms each of the single modal versions of the algorithm, demonstrating the effectiveness of the bimodal approach. The algorithm is implemented in a frame-by-frame manner with a low computational load, which makes it applicable for online applications.

The remainder of the paper is organized as follows. In Section II we formulate the problem. The construction of the low dimensional representation of the signals is described in Section III. The estimation of the speech presence indicator is described in Section IV, and experimental results demonstrating the performance of the proposed algorithm are presented in Section V.

II. Problem Formulation

Let $a[n]$ be a measured audio signal given by:

$$a[n] = a^s[n] + a^d[n] + a^t[n], \tag{1}$$

where $a^s[n]$, $a^d[n]$ and $a^t[n]$ are speech, background noise and transient interference, respectively.

The signal is processed in overlapping time frames of length M. Let $\mathbf{a}_i \in \mathbb{R}^M$ be the ith audio frame, and let $\mathcal{A} = \{\mathbf{a}_i\}_{i=1}^N$ be an audio data set of N time frames.

The video signal is assumed to comprise the region of the mouth, which is cropped out from a recorded front side video of a speaker. Note that the identification of the mouth region extends the scope of this paper. Nevertheless, we briefly describe in Section V the procedure performed in our experiments as a preprocessing stage. Let $\mathcal{V} = \{\mathbf{v}_i\}_{i=1}^N$ be the video data set comprising N consecutive video frames $\mathbf{v}_i \in \mathbb{R}^{W \times H}$, where W and H are the width and the height of the frame, respectively.

We assume that for each frame index i, $\mathbf{a}_i$ and $\mathbf{v}_i$ are aligned. Namely, both frames represent data captured by different sensors at the same time. The alignment is achieved by setting the length of the audio frame M to correspond the video frame rate.

Let $\mathcal{H}_0$ and $\mathcal{H}_1$ be two hypotheses denoting speech absence and presence, respectively. According to the hypotheses, we define a speech indicator as

$$\mathbb{1}_{\mathbf{s}}(i) = \left\{ \begin{array}{ll} 1, & i \in \mathcal{H}_1 \\ 0, & i \in \mathcal{H}_0 \end{array} \right\}. \tag{2}$$

The goal in this paper is to estimate $1_{\mathbf{s}}(i)$, i.e., to classify each frame as a speech or a non-speech frame.

We consider two audio-visual data sets. A training data set $\{\mathcal{A}^{tr}, \mathcal{V}^{tr}\}$ of size N^{tr} frames, and a test data set $\{\mathcal{A}^{te}, \mathcal{V}^{te}\}$ of size N^{te}. Each data set consists of both speech and non-speech intervals which are contaminated with noise and transients and are labeled according to the presence and absence of speech. The training data set is used to construct a low dimensional model of the data in a batch manner, and to train an estimator in a supervised manner, for the speech presence indicator. The test set is used for the evaluation of the proposed algorithm.

III. Low Dimensional Representation

The proposed algorithm is based on the representation of the audio-visual signal in a low dimensional domain where speech frames are separated from the non-speech frames. This representation is constructed using a manifold learning method which is applied in a high dimensional feature space of each modality.

Multiple notations for a frame are used throughout this paper. Frame i of the input signal is denoted by $\{\mathbf{a}_i, \mathbf{v}_i\}$, the corresponding high dimensional feature vector is denoted by $\{\tilde{\mathbf{a}}_i, \tilde{\mathbf{v}}_i\}$, and the low dimensional representation (obtained by manifold learning) is denoted by $\{\hat{\mathbf{a}}_i, \hat{\mathbf{v}}_i\}$.

A. Features Extraction

1) Audio Features: The proposed audio features are based on spectral representation of speech using MFCCs and the Short-Time Fourier Transform (STFT). These features were found to perform well for voice activity detection in challenging conditions, e.g. with a highly non-stationary noise [16]. Let $\mathbf{a}_i^{MFCC} \in \mathbb{R}^C$ be a column vector consisting of the MFCCs of frame $\mathbf{a}_i$, where C is the number of the coefficients. MFCCs are widely used in the field of speech recognition, since they successfully represent the spectrum of speech in a compact form using the perceptually meaningful Mel-frequency scale [28]. However, the MFCC representation of a speech frame may be similar to the representation of a non-speech frame comprising highly non-stationary noise. To improve the separation between the signal and the background noise, the MFCCs of each frame are weighted by a scalar which is based on noise estimation in the frame, such that a low value is assigned when merely the background noise is present [16].

Traditionally, the input signal is assumed to contain only speech and stationary noise. Thus, speech and noise are separated assuming that stationary noise components in the STFT domain are slowly varying with respect to speech [31], [32]. However, transients vary faster than speech, and hence, they are mistakenly identified as (non-stationary) speech [9], [10]. Therefore, the weights which are based on the noise estimation method presented in [32], in this case, only separate speech and transients from the stationary (or quasi-stationary) noise. Next, we describe the computation of such weights and explain how to reduce the effect of transients on the frame representation.

Let $A(i,j)$ be the STFT representation of the audio signal $a[n]$, where j is the frequency bin index, and i is the time frame index. Accordingly, the representation of (1) in the STFT domain is given by:

$$A(i,j) = A^s(i,j) + A^d(i,j) + A^t(i,j) \tag{3}$$

where $A^s(i,j)$, $A^d(i,j)$ and $A^t(i,j)$ are the STFTs of $a^s[n]$, $a^d[n]$ and $a^t[n]$, respectively. The corresponding variances are given by $\lambda_s(i,j) = E[|A^s(i,j)|^2]$, $\lambda_d(i,j) = E[|A^d(i,j)|^2]$, and $\lambda_t(i,j) = E[|A^t(i,j)|^2]$, where $E[\cdot]$ denotes an expected value.

Similarly to the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, let $\mathcal{H}_s$ and $\mathcal{H}_{ns}$ be hypotheses for a stationary signal (background noise) and a non-stationary signal (speech and transients), respectively. The corresponding conditional Probability Density Functions (PDF) are given by $p_r(\cdot;\mathcal{H}_s)$ and $p_r(\cdot;\mathcal{H}_{ns})$, respectively. The log likelihood ratio between the non-stationary signal and the noise in the jth frequency bin of frame i is defined by:

$$\Lambda_i(j) = \log\left(\frac{p_r\left(A(i,j);\mathcal{H}_{ns}\right)}{p_r\left(A(i,j);\mathcal{H}_s\right)}\right). \tag{4}$$

Let $\xi(i,j)$ be the a priori Non-Stationary Signal to Noise Ratio (NSSNR), which is given by:

$$\xi(i,j) = \frac{\lambda_s(i,j) + \lambda_t(i,j)}{\lambda_d(i,j)}, \tag{5}$$

and is estimated according to [31], and let $\gamma(i,j)$ be the a posteriori NSSNR:

$$\gamma(i,j) = \frac{|A(i,j)|^2}{\lambda_d(i,j)}. \tag{6}$$

The estimation of both the a priori and the a posteriori NSSNRs is based on the spectral variance of the background noise, $\lambda_d(i,j)$, which is estimated using the improved minima controlled recursive (IMCRA) method [32].

Assuming that the non-stationary signal and noise have a complex uncorrelated Gaussian distribution in the STFT domain (using only speech and stationary noise model) it can be shown that the log likelihood ratio in (4) is given by [33], [3]:

$$\Lambda_i(j) = \left(\frac{\gamma(i,j)\cdot\xi(i,j)}{1+\xi(i,j)} - \log\left(1+\xi(i,j)\right)\right). \tag{7}$$

Let Λ_i be the arithmetic mean of the log likelihood ratio over all frequency bins of frame $\mathbf{a}_i$. To reduce the dynamical range of Λ_i, which is large, since, for example, when the background noise is absent, $p_r(A(i,j);\mathcal{H}_s) \to 0$ in (4), the weight of each frame $\mathbf{a}_j$ is given by normalizing Λ_i as:

$$w_\Lambda(i) = 1 - e^{-\frac{\Lambda_i}{\epsilon}}, \tag{8}$$

where ϵ is a normalization parameter. Now, $w_\Lambda(i)$ receives values close to 1 when speech or transients are present and values close to 0 when only background noise is present in the frame.

The audio feature vector $\tilde{\mathbf{a}}_i$ of frame $\mathbf{a}_i$ is defined by collecting the weighted MFCCs of $2J^A+1$ adjacent frames:

$$\tilde{\mathbf{a}}_i = \begin{pmatrix} w_\Lambda\left(i-J^A\right)\cdot \mathbf{a}_{i-J^A}^{MFCC} \\ w_\Lambda\left(i-J^A+1\right)\cdot \mathbf{a}_{i-J^A+1}^{MFCC} \\ \vdots \\ w_\Lambda\left(i+J^A\right)\cdot \mathbf{a}_{i+J^A}^{MFCC} \end{pmatrix} \in \mathbb{R}^{(2J^A+1)\cdot C}. \quad (9)$$

It is worthwhile noting that Λ_i was previously used for voice activity detection in [3] and [4]. However, since Λ_i and $w_\Lambda(i)$ cannot exclusively indicate speech activity in the presence of transients, $w_\Lambda(i)$ is used in this paper as a feature that separates speech and transients from background noise. Transient effects are attenuated by taking into account several consecutive time frames. This reduces the influence of transients on a frame representation since the typical duration of a transient is assumed to be of the order of a single time frame. Thus, for $J^A \geq 1$, the coordinates of $\tilde{\mathbf{a}}_i$ in the presence of speech are expected to be more consistent than in the presence of transients. In practice, we assign relatively small values to J^A, since large J^A induces a high dimension of features and requires a large number of training samples to construct the low dimensional model.

Recall that the main advantage of the visual signal is its resistance to the acoustic environmental interferences including transients. Thus, to further improve the robustness to transients, we incorporate the visual signal.

2) Visual Features: The proposed visual features are based on motion vectors which were previously exploited for voice activity detection in [22] and [23], and are calculated using Lucas-Kanade method [34], [35]. Let $\mathbf{v}_i(x,y)$ denote the (x,y)th pixel of frame $\mathbf{v}_i$, and let $v_i(x,y)$ and $u_i(x,y)$ denote the horizontal and the vertical components of the motion vector (i.e. the velocity) of the corresponding pixel. We form a vector $\mathbf{v}_i^{MV} \in \mathbb{R}^{W\cdot H}$ by concatenating the absolute values of the velocities of each pixel, which are given by $\sqrt{[v_i(x,y)]^2+[u_i(x,y)]^2}$.

The video signal is characterized both by spatial information, i.e., the shape of the mouth, and by temporal information, i.e., the movement of the mouth. The shape of the mouth indicates on the presence of speech as the pronunciation of most of the phonemes is associated with open mouth [24]. However, the shape of the mouth cannot exclusively indicate on the presence of speech since, for example, the mouth can be completely closed in particular speech frames. Thus, temporal information may serve as a complement, i.e., the mouth movement may correctly indicate on the presence of speech. To capture both spatial and temporal information, motion vectors are calculated in a spatio-temporal neighborhood of each pixel in a frame.

Yet, small movements of the mouth may naturally occur during non-speech intervals, thereby wrongly indicating speech presence. To further improve the temporal characterization of speech, we collect $2J^V+1$ adjacent frames in time, and form the following feature vector $\tilde{\mathbf{v}}_i$:

$$\tilde{\mathbf{v}}_i = \left(\mathbf{v}_{i-J^V}^{MV}, \mathbf{v}_{i-J^V+1}^{MV}, \ldots, \mathbf{v}_{i+J^V}^{MV}\right)^T \in \mathbb{R}^{(2J^V+1)\cdot W\cdot H}. \quad (10)$$

Similarly to the parameter J^A in (9), the parameter J^V is set to a small value to confine the dimensions of the video features.

B. Diffusion Maps

Speech production is usually associated with a small set of physical constraints, e.g. the positions of lips, jaw and tongue, which control the shape of the vocal tract [36]. A common practice is to use a parametric model for the production of speech, where the parameters are related to the physical constraints [37]. Assuming a set of K such parameters implies that speech can be represented in a K dimensional space. Instead of assuming a rigid parametric model, in this work we exploit a data driven approach to learn a low dimensional representation of speech. Our main assumption is that the high dimensional feature vectors are not spread across the entire space, but rather lie on a manifold of a significantly lower dimension. In particular, since the features are specifically designed to emphasize the characteristics of speech, we assume that the manifold, i.e., the geometric structure of the features, is associated with the physical constraints of the production of speech. Therefore, the dimension of the manifold does not depend on the dimensions of the feature space which is dictated by the sensor (microphone or camera) that captures the signal.

In order to capture this low dimensional geometric structure, we use diffusion maps [29], which is a manifold learning method, that provides a parameterization of the data on the manifold through the embedding of the high dimensional feature vectors into a low dimensional space. In this work, diffusion maps is implemented by first constructing an empirical model of the manifold of the data using a training set, and then, the model of the manifold is extended to the test set in a frame-by-frame manner.

1) Construction of the Empirical Model Using the Training Set: Diffusion maps is applied similarly and separately to the feature vectors of each modality. Let $\tilde{\mathbf{f}}_i$ be the feature vector (audio or visual) of the ith frame. A pairwise similarity kernel function $k_\varepsilon(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_l)$ between the ith frame and the lth frame is defined as:

$$k_\varepsilon\left(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_l\right) = e^{-\frac{||\tilde{\mathbf{f}}_i - \tilde{\mathbf{f}}_l||^2}{\varepsilon}}, \quad (11)$$

where $\|\cdot\|$ is the L_2 norm and ε is the kernel bandwidth chosen according to [30]. Since the feature vectors are not uniformly distributed on the manifold, the kernel is normalized to provide a density invariant mapping [29]:

$$k_d\left(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_l\right) = \frac{k_\varepsilon\left(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_l\right)}{d\left(\tilde{\mathbf{f}}_i\right) d\left(\tilde{\mathbf{f}}_l\right)}. \quad (12)$$

where $d(\tilde{\mathbf{f}}_i)$ is the kernel normalization factor given by:

$$d\left(\tilde{\mathbf{f}}_i\right) = \sum_{\tilde{\mathbf{f}}_l \in \tilde{\mathcal{F}}^{tr}} k_\varepsilon\left(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_l\right), \quad (13)$$

where $\{\tilde{\mathcal{F}}^{tr}\}$ is a training set of N^{tr} (audio or video) feature vectors. Based on the kernel, a weighted symmetric graph is constructed, where each feature vector $\tilde{\mathbf{f}}_i$ is viewed as a node, and the weight of the edge connecting nodes i and l is given by $k_d(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_l)$. We now define a Markov chain on the graph by

normalizing the kernel once again. Let $\mathbf{M} \in \mathbb{R}^{N^{tr} \times N^{tr}}$ be a row stochastic Markov matrix, which is given by

$$\mathbf{M}_{i,l} = \frac{k_d\left(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_l\right)}{s\left(\tilde{\mathbf{f}}_l\right)}, \tag{14}$$

where

$$s\left(\tilde{\mathbf{f}}_i\right) = \sum_{\tilde{\mathbf{f}}_l \in \tilde{\mathcal{F}}^{tr}} k_d\left(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_l\right). \tag{15}$$

As a result, the nodes of the graph $\tilde{\mathbf{f}}_i \in \tilde{\mathcal{F}}^{tr}$ may be seen as the states of a Markov chain with the transition probability matrix $\mathbf{M}$. Finally, eigenvalue decomposition is applied to $\mathbf{M}$, yielding eigenvalues $\{\mu_k\}$, which are sorted in a descending order, and corresponding eigenvectors $\{\phi_k\}$. The eigenvalues of $\mathbf{M}$ are in the range of [0,1] due to the row normalization [29]. Moreover, $\mu_0 = 1$ and its associated eigenvector ϕ_0 is an all-ones vector. This constant eigenvector is ignored since it does not contain any information [38].

The K largest eigenvalues of $\mathbf{M}$ (excluding the trivial) and their corresponding K eigenvectors are used for the parameterization of the feature vectors on the manifold. We form a matrix $\mathbf{\Phi} \in \mathbb{R}^{N^{tr} \times K}$ whose columns consist of the eigenvectors and the eigenvalues of the transition probability matrix:

$$\mathbf{\Phi} \equiv (\mu_1\phi_1, \mu_2\phi_2, \ldots, \mu_K\phi_K). \tag{16}$$

From (16), the diffusion mapping of the feature vector $\tilde{\mathbf{f}}_i$ is given by the ith row of the matrix $\mathbf{\Phi}$:

$$\hat{\mathbf{f}}_i = (\mathbf{\Phi}_{i,1}, \mathbf{\Phi}_{i,2}, \ldots, \mathbf{\Phi}_{i,K}). \tag{17}$$

Thus, we obtain an embedding $\hat{\mathbf{f}}_i$ of each feature vector into a K dimensional Euclidean space. According to our assumption that there exists a low dimensional intrinsic structure of the data, the spectrum of the transition probability matrix (the eigenvalues) decays fast. Therefore, entries in (16) corresponding to small eigenvalues are negligible and K may be set to a small value, thereby providing significant dimensionality reduction. This property will be illustrated in Section V.

2) Online Processing of the Test Set: In Section III-B1, we construct a low dimensional empirical model for the training feature vectors in a batch manner. In this section, we show how to extend the model to new incoming frames. The extension is performed in a frame-by-frame manner similarly to the Nyström method [30], [39].

Let $\mathbf{f}_q$ and $\tilde{\mathbf{f}}_q$ be a new incoming test frame and its corresponding feature vector, respectively, and let $\mathbf{w} \in \mathbb{R}^{N^{tr}}$ be a weighting vector. The kth entry of the extended diffusion maps, ϕ'_k, is given by:

$$\phi'_k = \mathbf{w}^T \mathbf{\Phi}(:, k) \tag{18}$$

where $\mathbf{w}^T$ is the transpose of the weighting vector and $\mathbf{\Phi}(:, k)$ is the kth column of $\mathbf{\Phi}$. The extension may be seen as a weighted nearest neighbor interpolation, where $\mathbf{w}$ consists of the interpolation weights. The ith entry of the weighting vector represents the similarity between the incoming test frame $\mathbf{f}_q$ and the ith training frame $\mathbf{f}_i$. Thus, the closer the extended frame is to a particular training frame in the features domain, the higher the weight of the diffusion maps entry of the training frame is in the extension. Traditionally, when the Nyström method is used to extend eigenvectors of a matrix, in our case $\mathbf{M}$, $\mathbf{w}(i)$ is given by $\mathbf{M}_{i,q}$. However, due to the normalization applied in (12)–(15), $\mathbf{M}_{i,q}$ can not be properly calculated in a frame-by-frame manner. Therefore, the "true" interpolation weights are approximated by a Gaussian kernel with the following correction [30]. Let $\mathbf{k} \in \mathbb{R}^{N^{tr}}$ be a vector whose ith entry is given by a Gaussian kernel:

$$\mathbf{k}_i = e^{-\left(\frac{||\tilde{\mathbf{f}}_i - \tilde{\mathbf{f}}_q||}{\sigma}\right)^2}, \tag{19}$$

where σ is the kernel bandwidth. Similarly, let $\mathbf{K} \in \mathbb{R}^{N^{tr} \times N^{tr}}$ be a similarity matrix defined on the training set, whose (i, l)th entry is also given by the Gaussian kernel:

$$\mathbf{K}_{i,l} = e^{-\left(\frac{||\tilde{\mathbf{f}}_i - \tilde{\mathbf{f}}_l||}{\sigma}\right)^2}. \tag{20}$$

The weighting vector $\mathbf{w}$ is given by:

$$\mathbf{w} = \mathbf{K}^{-1}\mathbf{k}, \tag{21}$$

where the inverse matrix $\mathbf{K}^{-1}$ is used to correct the Gaussian weights. The weights are designed to provide a consistent extension of the diffusion maps. Namely, by substituting a training frame $\tilde{\mathbf{f}}_l$ instead of the test frame $\tilde{\mathbf{f}}_q$ into (19), the interpolation weight is degenerated to the Kronecker delta function, i.e., $\mathbf{w}_i = \delta_{i,l}$, and the extended value coincides with the true value, i.e., $\phi'_k = \mathbf{\Phi}(l, k)$.

Based on equation (18), the low dimensional representation of the test frame is given by:

$$\hat{\mathbf{f}}_q = (\phi'_1, \phi'_2, \ldots, \phi'_K). \tag{22}$$

This procedure is applied separately to each incoming test frame with a computational cost which is linear with the size of the training set, N^{tr}, making diffusion maps adequate for real time applications.

C. Diffusion Distance

Let $D(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_l)$ denote the diffusion distance between a pair of feature vectors $\tilde{\mathbf{f}}_i$ and $\tilde{\mathbf{f}}_l$, which is given by [38], [29]:

$$D(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_l) = \left(\sum_{q=1}^{N^{tr}} \frac{(\mathbf{M}_{i,q} - \mathbf{M}_{l,q})^2}{\nu_0(q)}\right)^{1/2} \tag{23}$$

where ν_0 is the unique stationary distribution of the Markov chain and is given by:

$$\nu_0(i) = \frac{s\left(\tilde{\mathbf{f}}_i\right)}{\sum_{\tilde{\mathbf{f}}_l \in \tilde{\mathcal{F}}^{tr}} s\left(\tilde{\mathbf{f}}_l\right)}. \tag{24}$$

The diffusion distance reflects the connectivity of the nodes (feature vectors) in the graph: Short distances are obtained for highly connected nodes due to high values of transition probabilities between the nodes [29]. The diffusion distance is known to be more robust to noise compared to the Euclidean distance,

since it integrates information from many features, whereas the Euclidean distance takes into account only two individual features. In addition, the diffusion distance is unit-less as it is calculated through transition probabilities. Therefore, it is suitable for merging data captured in different types of sensors.

When all the eigenvalues and the eigenvectors are used for the construction of diffusion maps in (17), i.e., $K = N^{tr}$, the L_2 distance in the diffusion maps domain equals the diffusion distance [29], [38]. Yet, even relatively small values of K (the dimension of the diffusion maps) provide an accurate approximation of the diffusion distance due to the fast spectrum decay:

$$D(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_l) \approx ||\hat{\mathbf{f}}_i - \hat{\mathbf{f}}_l||. \tag{25}$$

This approximation allows for an efficient computation of the diffusion distance from the embedding of the feature vectors.

The new representation is particularly suitable for estimating the speech presence indicator both because of the specific choice of the feature vectors that characterize speech, and because of the properties of the diffusion mapping. The later provides a low dimensional representation that captures the essence of the data and a good distance metric to compare embedded signal samples based on their intrinsic structure.

IV. Estimation of the Speech Presence Indicator

A. Unimodal Estimation of Speech Presence Indicator

Based on the low dimensional representation of the signals, we propose a continuous measure for voice activity, such that the clustering is achieved by comparing the measure to a threshold. This allows to control the tradeoff between false alarms and correct detections and the algorithm may be adjusted to the particular application at hand.

Let $P(\mathbf{f}_i)$ be the measure for voice activity in frame $\mathbf{f}_i$. $P(\mathbf{f}_i)$ comprises two components representing two different aspects of speech presence. The first, denoted by $P^S(\mathbf{f}_i)$, is supervised and relies on the diffusion distance between a test frame and the labeled training frames. The second, denoted by $P^{US}(\mathbf{f}_i)$, is derived using an unsupervised procedure and further exploits the dynamics of the signals.

$P^S(\mathbf{f}_i)$ is computed using a GMM procedure applied in the diffusion maps domain. Let $p_r(\hat{\mathbf{f}})$ be a Gaussian mixture PDF, given by:

$$p_r(\hat{\mathbf{f}}) = \sum_{r=1}^{R} \rho_r g\left(\hat{\mathbf{f}}|\zeta_r, \Sigma_r\right), \tag{26}$$

where R is the number of Gaussian components, $\rho_r, r = 1, 2, \ldots, R$ are the mixture weights that sum to one, and $g(\hat{\mathbf{f}}|\zeta_r, \Sigma_r)$ is the PDF of the rth Gaussian component, given by:

$$g\left(\hat{\mathbf{f}}|\zeta_r, \Sigma_r\right) = \frac{\exp\left(-\frac{1}{2}(\hat{\mathbf{f}} - \zeta_r)^T \Sigma_r^{-1}(\hat{\mathbf{f}} - \zeta_r)\right)}{(2\pi)^{K/2} \det\left(\Sigma_r\right)^{1/2}}, \tag{27}$$

where K is the dimension of $\hat{\mathbf{f}}$ and $\det(\Sigma_r)$ is the determinant of Σ_r. We assume two such GMMs, one for the speech absence hypothesis, $\mathcal{H}_0$, and the other for the speech presence hypothesis, $\mathcal{H}_1$. In order to estimate the parameters $\rho_r, \zeta_r, \Sigma_r, r = 1, 2, \ldots, R$ of each GMM, we use the training set, which is separated according to its labeling into two clusters, one for each hypothesis. The parameters of each GMM are estimated using the corresponding cluster by the expectation-maximization (EM) algorithm [40]. Let $p_r(\hat{\mathbf{f}}; \mathcal{H}_1)$ and $p_r(\hat{\mathbf{f}}; \mathcal{H}_0)$ be the Gaussian mixture PDFs of the speech and the non-speech clusters, respectively. Given a test frame $\mathbf{f}_i$, a bounded likelihood ratio between the conditional densities is calculated:

$$\Gamma_i = \min\left(\frac{p_r(\hat{\mathbf{f}}_i; \mathcal{H}_1)}{p_r(\hat{\mathbf{f}}_i; \mathcal{H}_0)}, \Gamma_{\max}\right), \tag{28}$$

where $\Gamma_{\max}$ is a constant value which is used to confine the dynamical range of the likelihood ratio. A likelihood ratio above this value, indicates voice activity with a high probability. In practice, we set $\Gamma_{\max} = 100$, and empirically found small influence on the performance of the algorithm for a wide range of values. According to (28), the closer $\hat{\mathbf{f}}_i$ is to the speech training cluster, the higher Γ_i level is.

The supervised measure for voice activity P^S is defined by:

$$P^S(\mathbf{f}_i) = \frac{1}{(2L^S + 1) \cdot \Gamma_{\max}} \sum_{l=-L^S}^{L^S} \Gamma_{i+l}. \tag{29}$$

Recall that in (9) and in (10), the use of the temporal neighborhood to characterize speech was limited to keep reasonable values of the dimensions of the features. In (29), the temporal neighborhood is exploited without these limitations, and Γ_i is averaged over $(2L^S + 1)$ consecutive frames to smooth the measure of voice activity and improve the estimation of speech presence indicator. For example in non-speech intervals, short term interruptions such as transients may provide instantaneous high values of Γ_i, yet, smoothing the measure over the temporal neighborhood provides correct low levels of voice activity.

Before we turn to describe the construction of the second voice activity measure, P^{US}, we note that integration of these measures requires them to be in the same range of values, which is set for simplicity to [0,1]. The factor $(2L^S + 1) \cdot \Gamma_{\max}$ in (29) properly keeps the values of P^S in this range. In this context, we remark that instead of using $\Gamma_{\max}$ to confine the dynamical range of Γ_i in (28), a different approach is to apply a log to the likelihood ratio. However, this approach does not confine the values of the measure to a finite range as is necessary for the integration between the measures. Another approach to get a finite value range for P^S is to use the class posterior of the GMM of the speech cluster. While this approach provides values in the range of [0,1], it was empirically found to provide inferior results compared to the performance of the proposed measure, P^S. One explanation to the inferior results is that this approach only exploits the training data of the speech cluster while the training data of the non-speech cluster is discarded.

The unsupervised activity measure P^{US} exploits the variability between consecutive frames in the test set in terms of diffusion distance and is defined by:

$$P^{US}(\mathbf{f}_i) = \frac{\min\left(\sum_{l=1}^{L^{US}} D(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_{i-l}), \sum_{l=1}^{L^{US}} D(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_{i+l})\right)}{L^{US} \cdot D_{\max}}, \tag{30}$$

where $D_{\max}$ is the maximal diffusion distance between a pair of frames in the training set, and $L^{US} \cdot D_{\max}$ is a normalization factor, which is used to keep the values of P^{US} (given a large enough training set), in the range of [0,1], similarly to the value range of P^S. In speech absence periods, audio frames tend to be similar to their adjacent frames as background noise varies slower than speech. A similar property is observed in video frames, as a slower mouth movement is assumed in speech absence periods compared to periods when speech is present. Accordingly, P^{US} is expected to provide lower levels of voice activity when speech is absent compared to when it is present, for both modalities. According to (30) the variability of the frames is measured in two non-overlapping windows: a causal window and an anti-causal window, both of size L^{US}. The min function is used to reduce false detection at the beginning and at the end of speech intervals. For example, correct low values of activity are received right after the end of a speech interval due to low values of variability in the anti-causal window despite high levels of variability in the causal window. Speech presence estimation based on P^{US} is viewed as an unsupervised procedure since the training data is used only for the construction of diffusion maps without its labeling.

The integrated activity measure of frame $\mathbf{f}_i$ is given by:

$$P(\mathbf{f}_i) = \frac{P^S(\mathbf{f}_i) + P^{US}(\mathbf{f}_i)}{2}. \tag{31}$$

The performance of the supervised measure P^S highly depends on the similarity of the tested signal to the training set. However, tested frames may be close to the wrong cluster in the training set due to differences, e.g., between speakers or acoustic conditions, in the tested and the training sets. For the unsupervised measure, the training set is utilized only for the construction of diffusion maps, and therefore it is more resistant to such differences. As a result, the integration between the measures provides an improved measure as we will show in Section V.

B. Bimodal Estimation of Speech Presence Indicator

Let $P(\mathbf{a}_i)$ and $P(\mathbf{v}_i)$ be the measures of voice activity from the audio and video signals, respectively. We compute the bimodal activity $P^B(\mathbf{a}_i, \mathbf{v}_i)$ according to:

$$P^B(\mathbf{a}_i, \mathbf{v}_i) = \alpha P(\mathbf{a}_i) + (1-\alpha)P(\mathbf{v}_i) \tag{32}$$

where α is in the range of [0,1] and controls the given weight to the two modalities. The setting of this parameter is application dependent. When the audio signal is relatively clean, α should be set close to 1. To quantify the quality of the audio signal, the estimate of the SNR in the audio signal may be used to adjust α over time. Further adaption of α may prevent failure of the algorithm in challenging real scenarios. For example, α may be set to 1 for frames where a speaker moves his head out of the frame, thereby making the video signal irrelevant.

In this work, for simplicity we set $\alpha = 0.5$. Combining the modalities this way was derived in [41] through a Bayesian model under restrictive assumptions that the modalities are statistically independent and that a posteriori probability of each modality remains close to the priors. Nevertheless, it was empirically found to outperform other fixed functions for combining the two modalities (such as a product, minimum, maximum, and median) due to better resistance to estimation errors. This simple combination empirically showed good results as illustrated in Section V. Adaptive setting of α will be addressed in a future work.

A different approach for combining the modalities can be achieved by concatenating the diffusion maps of each modality into a single super-vector [38], [30]. As a result, the speech presence indicator can be estimated in a unified diffusion maps domain, and $P(\cdot)$ in (31) represents a bimodal measure for speech presence. However, this approach does not allow to control the given weight to each one of the two modalities, as can be done by α in (32).

The proposed measure of voice activity $P^B(\mathbf{a}_i, \mathbf{v}_i)$ gets values in the range of [0,1], and hence, it can be viewed as a generalized a posteriori probability for speech. Finally, the estimate of the speech presence indicator is computed by comparing $P^B(\mathbf{a}_i, \mathbf{v}_i)$ to a threshold τ:

$$\hat{\mathbb{1}}_{\mathbf{s}}(i) = \begin{cases} 1 & ; \quad P^B(\mathbf{a}_i, \mathbf{v}_i) > \tau \\ 0 & ; \quad otherwise \end{cases}. \tag{33}$$

Future frames which are used in (9), (10), (29) and (30) induce a lag in online processing. The number of lagged frames, N^{lag}, is given by:

$$N^{lag} = \max\left(J^A, J^V\right) + \max(L^S, L^{US}) \tag{34}$$

The effect of the lag on real time processing is discussed in Section V.

Algorithm 1 summarizes the proposed VAD. For simplicity, the algorithm is presented under the assumption that N^{lag} future frames are available.

V. Experimental Results

A. Experimental Setup

The experimental setup simulates a video call made from a smartphone. The data set comprises 11 speakers reading aloud an article. During the experiments, the speakers make natural pauses every few sentences. As a results, the typical lengths of the recorded speech and non-speech intervals range from $100-300$ ms to $2-3$ s.

The video is recorded using a frontal camera of a smartphone (Samsung I9100, ~ 25 [fps], 640×480 resolution), providing front side videos of the speakers. The video is converted to gray scale to reduce the computational load. A bounding box of the mouth (110×90 pixels) is cropped out of the videos.

Although cropping the bounding box of the mouth extends the scope of this work, we shortly explain the procedure performed in our experiments as a preprocessing stage. The cropping is based on nostrils tracking, which are manually marked in the first frame, and are then searched in the following frames in a small area around their previous location. The search is performed under the assumption that the pixels, where the nostrils are located, have lower intensity values relatively to skin and lips pixels due to shading. Such a method for nostrils tracking was previously explored in [42].

The bounding box of the mouth is downsampled by a factor of 10 to reduce the computational load in the calculation of the motion vectors, and W and H are set to 11 and 9 (all the used

Algorithm 1 Audio-visual voice activity detection

procedure Training

Input: training data- $\{\mathbf{a}_i, \mathbf{v}_i\}_{i=1}^{N^{tr}}$

Output: diffusion maps- $\{\hat{\mathbf{a}}_i, \hat{\mathbf{v}}_i\}_{i=1}^{N^{tr}}$; estimate of the PDFs, $p_r(\hat{\mathbf{f}}; \mathcal{H}_0)$ and $p_r(\hat{\mathbf{f}}; \mathcal{H}_1)$, for the GMM

1: **for** $i = 1 : N^{tr}$ **do**
2: Calculate the feature vectors $\tilde{\mathbf{a}}_i$, $\tilde{\mathbf{v}}_i$
3: **end for**
4: Do for each modality separately:
5: Calculate the transition probability matrix $\mathbf{M}$ using (11)–(15)
6: Apply eigenvalue decomposition on $\mathbf{M}$ and obtain the eigenvalues $\{\mu_k\}$ and the eigenvectors $\{\phi_k\}$
7: Build diffusion maps $\hat{\mathbf{f}}_i, i \in [1, 2], \ldots, N^{tr}]$ ($\hat{\mathbf{a}}_i$ for audio and $\hat{\mathbf{v}}_i$ for video) using (16) and (17)
8: Train the GMMs using the labeling and estimate $p_r(\hat{\mathbf{f}}; \mathcal{H}_0)$ and $p_r(\hat{\mathbf{f}}; \mathcal{H}_1)$

end procedure

procedure Test

Input: test data- $\{\mathbf{a}_i, \mathbf{v}_i\}_{i=1}^{N^{te}}$

Output: speech presence indicator estimate- $\{\hat{\mathbb{1}}_{\mathbf{s}}(i)\}_{i=1}^{N^{te}}$

1: Get a new frame $\mathbf{a}_i$, $\mathbf{v}_i$
2: Calculate the feature vectors $\tilde{\mathbf{a}}_i$, $\tilde{\mathbf{v}}_i$
3: Extend the diffusion maps $\hat{\mathbf{a}}_i$, $\hat{\mathbf{v}}_i$ using (18) and (22)
4: Do for each modality separately:
5: Calculate the first voice activity measure $P^S(\mathbf{f}_i)$ using the PDFs of the GMM according to (28) and (29)
6: Calculate the second voice activity measure $P^{US}(\mathbf{f}_i)$ according to (30)
7: Integrate the measures: $P(\mathbf{f}_i) = \frac{P^S(\mathbf{f}_i) + P^{US}(\mathbf{f}_i)}{2}$
8: Merge the modalities: $P^B(\mathbf{a}_i, \mathbf{v}_i) = \alpha P(\mathbf{a}_i) + (1 - \alpha) P(\mathbf{v}_i)$
9: **if** $P^B(\mathbf{a}_i, \mathbf{v}_i) > \tau$ **then**
10: Decide $\mathcal{H}_1$
11: **else**
12: Decide $\mathcal{H}_0$
13: **end if**
14: Go back to 1

end procedure

parameters values are presented in Table I). An example of a speech frame image and an illustration of the corresponding motion vectors are presented in Fig. 1, demonstrating that motion vectors capture the shape of the mouth and its movement with respect to the previous frame.

The audio is recorded by the microphone of the smartphone and is processed in 8 [kHz] (higher processing rates have shown no advantage). The recordings are performed in a quiet room (estimated audio SNR of $\sim$ 25 [dB]) and are regarded as a clean audio signal. The audio signal is processed using short time frames of length $M = 634$ with 50% overlap. Such a configuration aligns the rates of the audio and video signals.

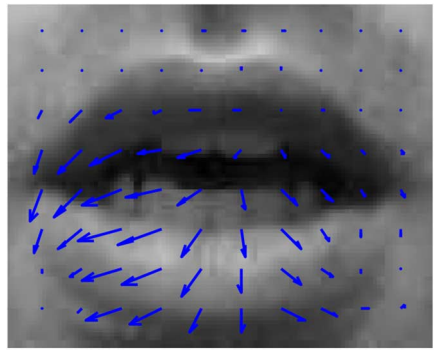


Fig. 1. Motion vectors in a speech interval and the corresponding video frame.

TABLE I
ALGORITHM PARAMETERS

Type	Parameter	
Audio features	Frame length	$M = 634$
	Normalization parameter	$\epsilon = 3$
	Number of MFCCs	$C = 24$
	Temporal characterization	J_A=1
Video features	Cropped frame size	$W = 11$, $H = 9$
	Temporal characterization	$J_V = 1$
Diffusion maps	Extension kernel bandwidth	$\sigma = 50$
	Number of eigenvalues	$K = 4$
Estimation	Number of Gaussians for $\mathcal{H}_1$	5
	Number of Gaussians for $\mathcal{H}_0$	5
	Max. likelihood ratio value	$\Gamma_{\max} = 100$
	Time window in P^S	$L^S = 9$
	Time window in P^{US}	$L^{US} = 9$

The training data set is created by collecting 30 [sec] long data sequences of 6 speakers (the total training data set is 180 [sec], 4542 frames). We empirically find that a 180 [sec] long signal both contains sufficient amount of training data and the eigenvalue decomposition can be efficiently applied to $\mathbf{M}$ (using Intel Core i5-2500 CPU and 4 GB RAM). To make the calculation more efficient, $\mathbf{M}$ may be processed in blocks similarly to [16].

The algorithm is trained for challenging acoustic environments: various background noise types, which include white Gaussian noise, musical instruments noise, and babble noise, and various transient interferences, such as metronome, keyboard typing, and hammering, taken from [43], are added to the training audio signal of each speaker. The algorithm is trained for 0 and 5 dB SNR values, and the transients are normalized to have maximal amplitude twice larger than the maximal amplitude of speech. The training data of each speaker contains all possible combinations of background noise and transients. This training setup extends the setup in [16], where merely a single background noise and a single transient type were used for the training in each experiment.

The algorithm is tested using 60 s long data sequences of each of the 11 speakers. To prevent over fitting, for each tested speaker the algorithm is retrained with training data which do not contain the tested speaker.

B. *Qualitative Evaluation*

The 20 largest eigenvalues of the audio and video sets, $\{\mu_k\}$ in (16), are plotted in a decreasing order in Fig. 2, demonstrating fast decay of the spectrum. A spectral gap can be seen between

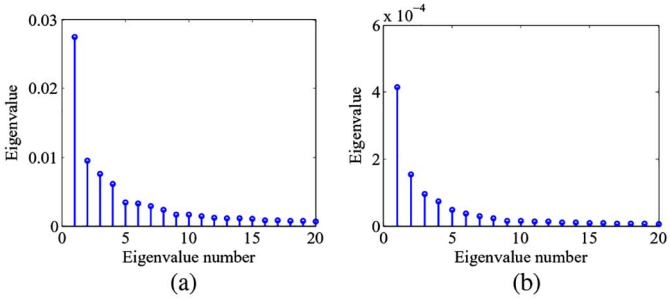


Fig. 2. Twenty largest eigenvalues of transition probability matrix, (a) audio, (b) video.

the fourth and fifth eigenvalues for the audio signal. This gap, heuristically implies that the intrinsic dimension of the signal is 4 [44]. Consequently, we set the dimension of diffusion maps to $K = 4$ in (17) for each modality. The choice of 8 parameters (4 for each modality) for representing a frame yields a significant dimensionality reduction of the data, and hence, allows for low computational complexity of the estimation procedure. In addition, these parameters capture just the essence of the data without noise and other nuisance factors, thereby allowing accurate detection of voice activity. The difference in the visual data representation from most of the previous studies, such as [24] and [27], is that in this work the parameters representing the mouth movement are obtained implicitly in a data-driven manner and are not defined in advance.

The measures of voice activity P^S in (29) and P^{US} in (30) are calculated using $L^S = L^{US} = 9$ past and future frames. This configuration induces a lag of $N^{lag} = 10$ frames in (34), which is $\sim$ 400 ms for $\sim$ 25 fps frame rate. Our experiments showed that lower values of lag may be set at the expense of a small degradation of the performance. In addition, for the evaluation of P^S we set the number of Gaussians to 5 to model both the speech and the non-speech clusters. To set this value, we empirically evaluated the performance of the algorithm using different number of Gaussians to model the speech and the non-speech clusters. We empirically found that the use of a small number of Gaussians for each one of the clusters provides good classification results, which could be explained by the compact representation of the data using diffusion maps. Our experiments showed no advantage of a higher number of Gaussians neither for the speech nor for the non-speech clusters. The number of Gaussians for the non-speech cluster is similar to that of the speech cluster due to the challenging conditions during non-speech intervals which include transients for the audio signal and non-speech lips movements for the video signal.

An example of the obtained voice activity detection is shown in Fig. 3. The input signal (black solid line) in this example is contaminated with a 10 dB babble noise and keyboard typing transients. Despite the frequent appearance of the transients, we observe an accurate speech presence indicator estimation when compared to the marked ground truth. We note that in this example, the threshold τ in (33) is empirically chosen to provide best estimation results. Although it is not in the scope of this work, in practice, the threshold may be set using the training data by evaluating the performance of the algorithm using a validation set.

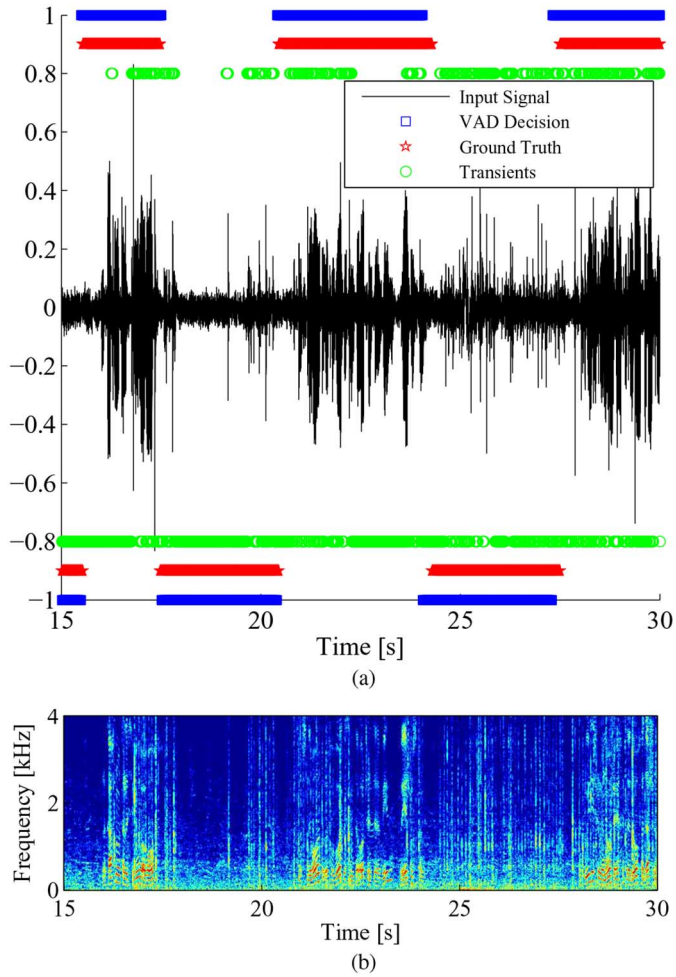


Fig. 3. Qualitative assessment of AV-VAD, with babble noise with 10 dB SNR and keyboard typing transient interferences. (a) Input signal- black solid line, ground truth - red stars graph, speech presence estimation using the proposed method - blue squares graph, the locations of the transients - green circles graph. (b) A spectrogram of the input signal.

C. Performance Evaluation Measure

For voice activity detection, the ground truth may be application dependent. For example, in speech recognition applications, isolated phonemes may be useful, and hence, the voice detection should ideally have a fine resolution in the order of few tens of milliseconds. On the other hand, in videoconferencing systems, where a central processing unit switches between cameras according to a dominant speaker [2], frequent switching between speakers should be avoided, and hence, coarser voice activity detection is required.

In addition, the ground truth depends on the modality (audio or video). Speech onsets, for example, may be accompanied with air aspiration, which is helpful for visual speech perception, and therefore, is considered as speech for the video signal but not for the audio signal. Another example is voiced phonemes. While the ends of voice phonemes are audible, they are not visual due to the lack of a mouth movement.

Therefore, we extend the definition of the speech indicator in (2). Similarly to the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, let $\mathcal{H}_0^a$ and $\mathcal{H}_1^a$ be the hypotheses for speech absence and speech presence in the

audio signal, respectively. Accordingly, let $\mathbb{1}_{\mathbf{s}}(\mathbf{a}_i)$ be an audio speech indicator, given by:

$$\mathbb{1}_{\mathbf{s}}(\mathbf{a}_i) = \left\{ \begin{array}{lcl} 1 & ; & \mathbf{a}_i \in \mathcal{H}_1^a \\ 0 & ; & \mathbf{a}_i \in \mathcal{H}_0^a \end{array} \right\}. \tag{35}$$

The audio speech indicator is manually marked using a spectrogram of a clean speech signal with a resolution of 100 [msec]. Similarly, let $\mathcal{H}_0^v$ and $\mathcal{H}_1^v$ be the hypotheses for speech absence and speech presence in the video signal, respectively. Accordingly, let $\mathbb{1}_{\mathbf{s}}(\mathbf{v}_i)$ be a video speech indicator given by:

$$\mathbb{1}_{\mathbf{s}}(\mathbf{v}_i) = \left\{ \begin{array}{lcl} 1 & ; & \mathbf{v}_i \in \mathcal{H}_1^v \\ 0 & ; & \mathbf{v}_i \in \mathcal{H}_0^v \end{array} \right\}. \tag{36}$$

The video speech indicator is manually marked as speech present when mouth moves during speech intervals (a natural mouth movement when speech is absent is neglected). The unified speech indicator $\mathbb{1}_{\mathbf{s}}(i)$ which is defined in (2) is given by:

$$\mathbb{1}_{\mathbf{s}}(i) = \mathbb{1}_{\mathbf{s}}(\mathbf{a}_i) \vee \mathbb{1}_{\mathbf{s}}(\mathbf{v}_i) \tag{37}$$

where $\vee$ is an "or" function. This setting may be adequate for applications such as audio-visual speech coding, where speech should be counted for each one of the sensors.

The quantitative performance is evaluated in three experiments. In the first and second experiments, unimodal versions of the proposed algorithm are compared to the state of the art and recently presented VADs. To evaluate the performance of the unimodal versions, $P^B(\mathbf{a}_i, \mathbf{v}_i)$ in (32) is replaced with the single modality activity measure P defined in (31): $P(\mathbf{a}_i)$ for audio and $P(\mathbf{v}_i)$ for video. For these experiments, the ground truth is given by (35) and (36) for evaluation based on only the audio signal and only the video signal, respectively. In the third experiment, the proposed AV-VAD is compared to the single modality versions using the audio-visual ground truth given by (37).

D. VAD Evaluation

The performance of the proposed algorithm based on the audio signal is compared to the methods presented in [3], [4], [5], [14] and [16]. Similarly to the proposed algorithm, in the algorithms presented in [4] and [16], the likelihood ratio is calculated in past and future frames. This allows for activity level smoothing and is more adequate for the ground truth setting in this work. However, in the VAD presented in [3], [5] and [14] the likelihood ratio is calculated for a single frame, which makes the detection less adequate to this application. To make a fair comparison, we smooth the VADs in [3], [5] and [14] with a median filter of length 19, which significantly improves their performance.

The proposed algorithm based on the video is compared to the method presented in [24]. Unlike the proposed algorithm, the VAD presented in [24] is not designed to perform in real time, as the parameters of the noise statistics in [24] are estimated in a batch manner. For simplicity of the implementation, these parameters are estimated using the ground truth of the test data set. We remark that estimation using the training set was also performed as suggested in [23]. Although it allows real time processing, our experiments show significant degradation of the

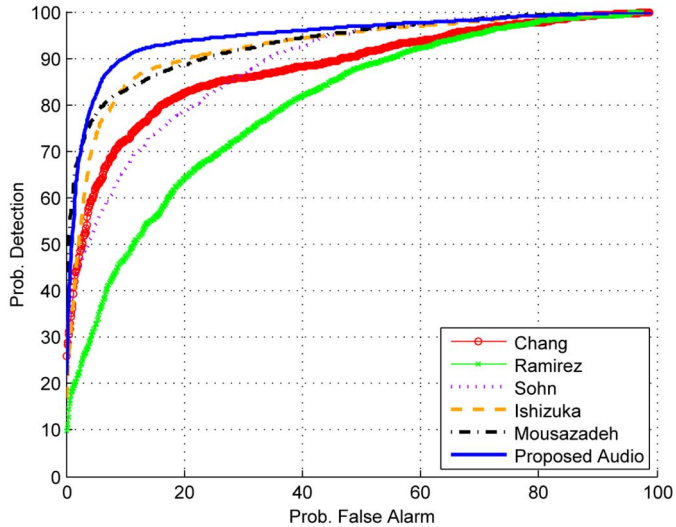


Fig. 4. Audio algorithms. Probability of detection vs probability of false alarm. Test for babble noise with 10 dB SNR and keyboard typing transient interferences.

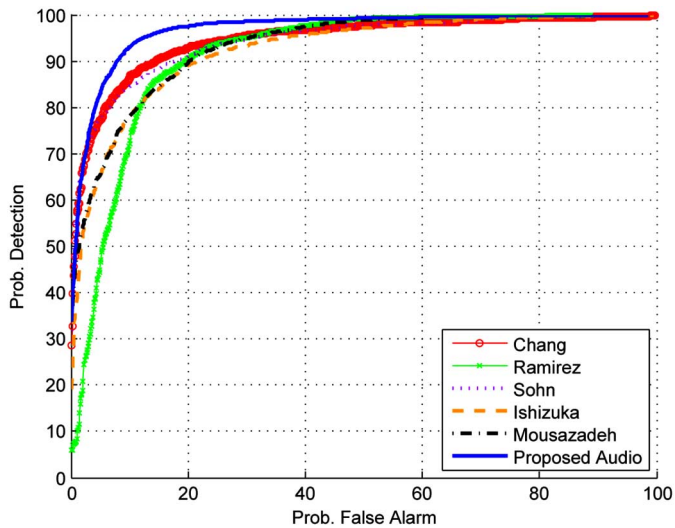


Fig. 5. Audio algorithms. Probability of detection vs probability of false alarm. Test for musical instruments noise with 10 dB SNR and hammering transient interferences.

performance of the competing algorithm, and hence, these results are not presented in the figures. In addition, we also evaluated the algorithm presented in [27]. We found that the procedure that separates the lips from the skin performed poorly on some of the videos in our data set, probably due to the lightening conditions. As a result, the overall performance of the algorithm was not comparable to the other two algorithms, and hence, is not presented in the figures.

Figs. 4–8 present the obtained Receiver Operating Characteristic (ROC) curves, which are generated by spanning the threshold over all possible activity values. The maximal performance of each method is presented in Tables II and III and is obtained using the threshold which provides the best results in terms of correct detection rate plus correct rejection rate.

In Figs. 4 and 5, we present the results of the evaluation of the algorithms based on the audio signal, where the curves marked by "Chang," "Ramirez," "Sohn," "Ishizuka" and "Mousazadeh" relate to the methods presented in [5], [4], [3], [14] and [16],

TABLE II
AUDIO ALGORITHMS RESULTS IN TERMS OF CORRECT DETECTION RATE PLUS CORRECT REJECTION RATE IN PERCENTS

	Babble 10 dB SNR Keyboard	Musical 10 dB SNR Hammering	Colored 5 dB SNR Hammering	Musical 0 dB SNR Keyboard	Babble 15 dB SNR Scissors
Chang	82	88.4	87	74.6	88.2
Ramirez	72.2	85.5	82.6	75.1	70.6
Sohn	79.4	87.5	84	80.7	88.4
Ishizuka	87.3	84.7	88.9	64.7	91.5
Mousazadeh	86.7	84.7	84.1	87.6	91
Proposed Audio	**90.1**	**91.5**	**90.5**	**88.7**	**92.4**

TABLE III
AUDIO-VISUAL ALGORITHMS RESULTS IN TERMS OF CORRECT DETECTION RATE PLUS CORRECT REJECTION RATE IN PERCENTS

	Babble 10 dB SNR Keyboard	Musical (10 dB SNR) Hammering	Colored 5 dB SNR Hammering	Musical 0 dB SNR Keyboard	Babble 15 dB SNR Scissors
Tamura	73.6	83.8	83.9	73.8	81.2
Proposed Audio	87.7	89.9	87.8	86.5	90.2
Proposed Video	89.6	89.6	89.6	89.6	89.6
Proposed P^{US}	89.3	89.8	89.6	89.5	85.8
Proposed P^{S}	92.3	94.2	92.1	92	93.8
Proposed AV	**92.9**	**94.5**	**92.8**	**92.9**	**94.6**

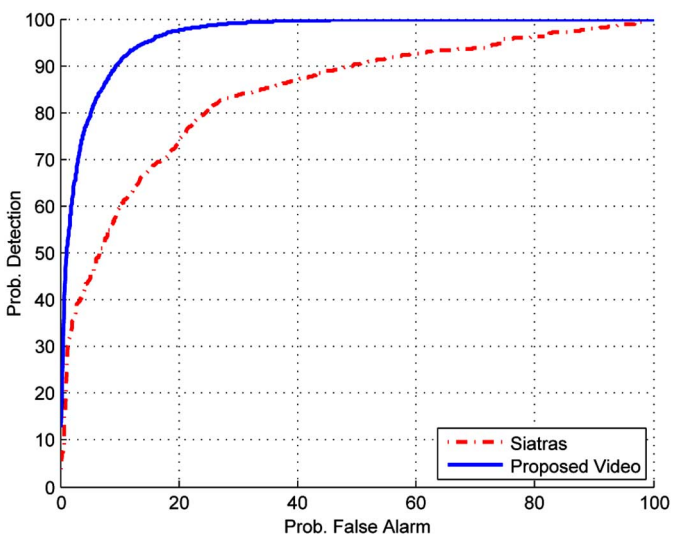


Fig. 6. Video algorithms. Probability of detection vs probability of false alarm. Best results in terms of correct detection rate plus correct rejection rate: Siatras-77.5%, Proposed Audio-90.6%.

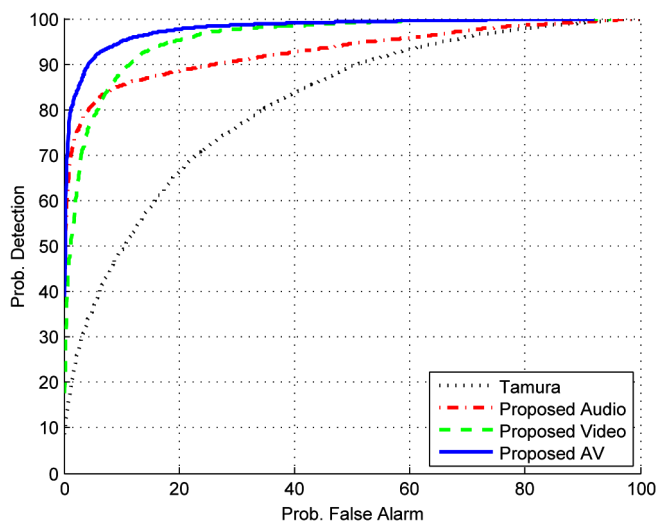


Fig. 7. Audio, video and audio-visual algorithms. Probability of detection vs probability of false alarm. Test for babble noise with 10 dB SNR and keyboard transient interferences.

respectively. In Fig. 4, the algorithm is tested for babble noise with 10 dB SNR and keyboard typing transients and in Fig. 5 for musical instruments noise with 10 dB SNR and hammering transients.

It can be seen in Fig. 4 that the methods "Chang," "Ramirez" and "Sohn," where the likelihood ratio is estimated assuming slow variation of the noise spectrum, provide inferior results in the presence of keyboard transient which is estimated as speech due to the fast variation of its spectrum. In Fig. 5, the performance of the method "Ishizuka" are degraded since musical instruments noise may be mistakenly detected as speech due to its periodic nature in the frequency domain, and the performance of the method "Mousazadeh" are degraded since the spectral clustering method poorly separates between the speech and the non-speech clusters. It can be seen in both figures that the proposed algorithm outperforms the competing audio VADs for the

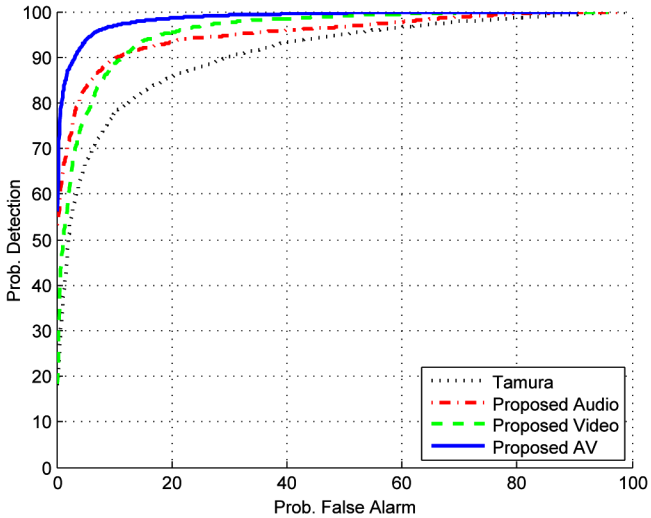


Fig. 8. Audio, video and audio-visual algorithms. Probability of detection vs probability of false alarm. Test for musical instruments noise with 10 dB SNR and hammering transient interferences.

different types of background noises and transients. We emphasize that the presented results of the proposed algorithm are achieved with a single training set consisting of the different types of background noises and transients. Namely, the type of the tested background noise and transient are not known in advance. The improved performance of the proposed algorithm based on the audio signal are further demonstrated in Table II for different types of background noises and transients and different SNR values.

The results of the algorithms based on the video signal are presented in Fig. 6, where the curve marked by "Siatras" relates to the method presented in [24]. It can be seen that the proposed algorithm outperforms the VAD in [24] for all possible values of false alarm rates. In addition, a high slope of the ROC curve of the proposed VAD is observed for low false alarm rates, thereby providing fast convergence to high detection rates.

In the third experiment, we evaluate the performance of the proposed AV-VAD and compare it to the performance of the algorithm using single modalities. In addition, we compare the proposed algorithm to the AV-VAD presented in [25].

The performances of the algorithms are presented in Figs. 7, 8 and in Table III. In Fig. 7, the algorithms are tested for babble noise with 10 dB SNR and keyboard typing transients, and in Fig. 8 the algorithms are tested for musical instruments noise with 10 dB SNR and hammering transients. It can be seen in both figures that the proposed algorithm outperforms the method presented in [25] which is marked in the plots by "Tamura." While the modalities in the proposed method are merged in (32) similarly to the merging scheme in [25], the main difference between the methods is that in this study, the representation of each modality is learned from the data and therefore, allows improved separation between the speech and the non-speech frames. In addition, due to the noisy acoustic conditions, the audio and the video versions of the proposed algorithm provide comparable results. In particular, they complement each other such that the audio version better performs for low values of false alarm, and the video version is better for the high values. The proposed bimodal algorithm embodies the advantages of each of the modalities, and provides best performance for each false alarm value. In Table III we also evaluate the performance of versions of the proposed algorithm where the speech indicator is estimated based solely on the supervised activity measure, P^S presented in (29), or the unsupervised activity measures, P^{US} presented in (30). It can be seen in the table that these measures complement each other. In particular for babble noise with 15 dB SNR and scissors transient, the unsupervised measure provides relatively low results since this type of transient is characterized by higher variability over time, compared to the other transients, which leads to false alarms. Yet, this transient is successfully separated from the speech signal in the diffusion maps domain and the supervised measure allows for an accurate classification. The proposed algorithm, which is based on the integrated measure in (31), performs better than the versions of the algorithm which are based solely on P^S or P^{US}, and provides the best performance for the different types of noises and transients.

VI. Conclusions

We have presented an algorithm for audio-visual voice activity detection. The algorithm is based on a low dimensional representation of the audio and the video signals which is constructed by applying diffusion mapping to features which are specifically designed to separate speech from non-speech frames. This representation of the signals is robust to noise, and facilitates a measure for voice activity that takes into account both training labeled data as well as the temporal variability of the signals. In addition, since diffusion maps are unit-less, the low dimensional representation is particularly suitable for processing data captured in different types of sensors. Experimental results have demonstrated that the proposed VAD based merely on the audio or the video signal outperforms state-of-the-art VADs. In addition, it has been shown that the proposed algorithm based on both the audio and the video outperforms each of the unimodal VADs and provides accurate voice activity detection in adverse noisy environments.

In the present study, equal weights are assigned to the two modalities in the merging scheme. In future research, we intend to develop adaptive merging schemes, which incorporate estimates of the quality of the (audio and video) signals. Another future research direction is further improving the separation between speech and transients. This may be achieved by including estimates of the transients (e.g., as proposed in [9]).

Acknowledgment

The authors thank the associate editor and the anonymous reviewers for their constructive comments and useful suggestions. The authors would like to thank Prof. Yosi Keller from Bar-Ilan University for providing the software implementation of diffusion maps.

References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, 2010.

[2] I. Volfin and I. Cohen, "Dominant speaker identification for multipoint videoconferencing," *Comput. Speech Lang.*, vol. 27, no. 4, pp. 895–910, 2013.

[3] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
[4] J. Ramìrez, J. Segura, C. Benìtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, vol. 42, no. 3, pp. 271–287, 2004.
[5] J. H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, Jun. 2006.
[6] J. Ramìrez, P. Yélamos, J. M. Górriz, and J. C. Segura, "SVM-based speech endpoint detection using contextual speech features," *Electron. Lett.*, vol. 42, no. 7, pp. 426–428, 2006.
[7] J. Ramìrez, J. C. Segura, and J. M. Górriz, "Revised contextual LRT for voice activity detection," in *Proc. 32nd IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2007, vol. 4, p. IV-801.
[8] M. Fujimoto and K. Ishizuka, "Noise robust voice activity detection based on switching Kalman filter," *IEICE Trans. Inf. Syst.*, vol. 91, no. 3, pp. 467–477, 2008.
[9] A. Hirszhorn, D. Dov, R. Talmon, and I. Cohen, "Transient interference suppression in speech signals based on the OM-LSA algorithm," in *Proc. Int. Workshop Acoust. Signal Enhance. (IWAENC)*, 2012, pp. 1–4.
[10] R. Talmon, I. Cohen, and S. Gannot, "Single-channel transient interference suppression with diffusion maps," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 1, pp. 132–144, Jan. 2013.
[11] D. Dov and I. Cohen, "Voice activity detection in presence of transients using the scattering transform," in *Proc. IEEE 28th Conv. Elect. Electron. Eng. Israel (IEEEI)*, 2014, pp. 1–5.
[12] R. Talmon, I. Cohen, and S. Gannot, "Transient noise reduction using nonlocal diffusion filters," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1584–1599, Aug. 2011.
[13] R. Talmon, I. Cohen, S. Gannot, and R. Coifman, "Supervised graph-based processing for sequential transient interference suppression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 9, pp. 2528–2538, Nov. 2012.
[14] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, "Noise robust voice activity detection based on periodic to aperiodic component ratio," *Speech Commun.*, vol. 52, no. 1, pp. 41–60, 2010.
[15] R. J. Weiss and T. T. Kristjansson, "DySANA: Dynamic speech and noise adaptation for voice activity detection," in *Proc. Annual Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2008, pp. 127–130.
[16] S. Mousazadeh and I. Cohen, "Voice activity detection in presence of transient noise using spectral clustering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 6, pp. 1261–1271, Jun. 2013.
[17] D. Sodoyer, B. Rivet, L. Girin, J. Schwartz, and C. Jutten, "An analysis of visual speech information applied to voice activity detection," in *Proc. 31st IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2006, vol. 1, pp. I–I.
[18] B. Rivet, L. Girin, and C. Jutten, "Visual voice activity detection as a help for speech source separation from convolutive mixtures," *Speech Commun.*, vol. 49, no. 7, pp. 667–677, 2007.
[19] A. Aubrey, B. Rivet, Y. Hicks, L. Girin, J. Chambers, and C. Jutten, "Two novel visual voice activity detectors based on appearance models and retinal filltering," in *Proc. 15th Eur. Signal Process. Conf. (EUSIPCO)*, 2007.
[20] D. Scott, C. Jung, J. Bins, A. Said, and A. Kalker, "Video based VAD using adaptive color information," in *Proc. 11th IEEE Int. Symp. Multimedia (ISM)*, 2009, pp. 80–87.
[21] C. B. O. Lopes, A. L. Goncalves, J. Scharcanski, and C. R. Jung, "Color-based lips extraction applied to voice activity detection," in *Proc. 8th IEEE Int. Conf. Image Process. (ICIP)*, 2011, pp. 1057–1060.
[22] A. J. Aubrey, Y. A. Hicks, and J. A. Chambers, "Visual voice activity detection with optical flow," *IET Image Process.*, vol. 4, no. 6, pp. 463–472, 2010.
[23] P. Tiawongsombat, M. H. Jeong, J. S. Yun, B. J. You, and S. R. Oh, "Robust visual speakingness detection using bi-level HMM," *Pattern Recogn.*, vol. 45, no. 2, pp. 783–793, 2012.
[24] S. Siatras, N. Nikolaidis, M. Krinidis, and I. Pitas, "Visual lip activity detection and speaker detection using mouth region intensities," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 1, pp. 133–137, Jan. 2009.
[25] S. Tamura, M. Ishikawa, T. Hashiba, S. T. , and S. Hayamizu, "A robust audio-visual speech recognition using audio-visual voice activity detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2010, pp. 2694–2697.
[26] T. Yoshida and K. Nakadai, "Two-layered audio-visual integration in voice activity detection and automatic speech recognition for robots," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2010, pp. 2702–2705.
[27] V. P. Minotto, C. B. O. Lopes, J. Scharcanski, C. R. Jung, and B. Lee, "Audiovisual voice activity detection based on microphone arrays and color information," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 1, pp. 147–156, Feb. 2013.
[28] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. 1st Int. Conf. Music Inf. Retrieval (ISMIR)*, 2000.
[29] R. R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 21, no. 1, pp. 5–30, 2006.
[30] Y. Keller, R. R. Coifman, S. Lafon, and S. W. Zucker, "Audio-visual group recognition using diffusion maps," *IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 403–413, Jan. 2010.
[31] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, 2001.
[32] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.
[33] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[34] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *Int. J. Comput. Vis.*, vol. 12, no. 1, pp. 43–77, 1994.
[35] A. Bruhn, J. Weickert, and C. Schnörr, "Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods," *Int. J. Comput. Vis.*, vol. 61, no. 3, pp. 211–231, 2005.
[36] B. H. Story, "A parametric model of the vocal tract area function for vowel and consonant simulation," *J. Acoust. Soc. Amer.*, vol. 117, no. 5, pp. 3231–3254, 2005.
[37] S. Mousazadeh and I. Cohen, "AR-GARCH in presence of noise: Parameter estimation and its application to voice activity detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 916–926, May 2011.
[38] S. Lafon, Y. Keller, and R. R. Coifman, "Data fusion and multicue data matching by diffusion maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1784–1797, Nov. 2006.
[39] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nyström method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 214–225, Feb. 2004.
[40] J. A. Bilmes *et al.*, "A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," *Int. Comput. Sci. Inst.*, vol. 4, no. 510, p. 126, 1998.
[41] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.
[42] P. S. Aleksic and A. K. Katsaggelos, "Audio-visual biometrics," *Proc. IEEE*, vol. 94, no. 11, pp. 2025–2044, Nov. 2006.
[43] [Online]. Available: http://www.freesound.org
[44] R. Talmon, I. Cohen, S. Gannot, and R. Coifman, "Diffusion maps for signal processing: A deeper look at manifold-learning techniques based on kernels and graphs," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 75–86, Jul. 2013.

David Dov received the B.Sc. (Summa Cum Laude) and M.Sc. degrees in electrical engineering from the Technion-Israel Institute of Technology, Haifa, Israel, in 2012 and 2014, respectively. He is currently pursuing the Ph.D. degree in electrical engineering at The Technion-Israel Institute of Technology, Haifa, Israel.

From 2010 to 2012, he worked in the field of Microelectronics in RAFAEL Advanced Defense Systems LTD. Since 2012, he has been a Teaching Assistant and a Project Supervisor with the Signal and Image Processing Lab (SIPL), Electrical Engineering Department, Technion. His research interests include speech processing, geometric methods for data analysis, multi-sensors signal processing, and multimedia.

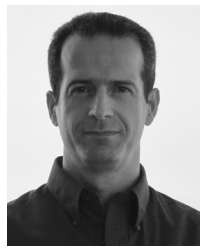
Mr. Dov is the recipient of the Meyer Fellowship and the Cipers Award for 2012, the Excellent Project Award for 2012, the Excellence in Teaching Award for outstanding teaching assistants for 2013, and the Jacobs Fellowship for 2014.

Ronen Talmon received the B.A. degree (Cum Laude) in mathematics and computer science from the Open University in 2005, and the Ph.D. degree in electrical engineering from the Technion in 2011. He is an Assistant Professor of electrical engineering at the Technion Israel Institute of Technology, Haifa, Israel.

From 2000 to 2005, he was a Software Developer and Researcher at a technological unit of the Israeli Defense Forces. From 2005 to 2011, he was a Teaching Assistant at the Department of Electrical Engineering, Technion. From 2011 to 2014, he was a Gibbs Assistant Professor in the Mathematics Department at Yale University, New Haven, CT. In 2014, he joined the Department of Electrical Engineering of the Technion.

His research interests are statistical signal processing, analysis and modeling of signals, speech enhancement, biomedical signal processing, applied harmonic analysis, and diffusion geometry.

Dr. Talmon is the recipient of the Irwin and Joan Jacobs Fellowship, the Andrew and Erna Fince Viterbi Fellowship, and the Horev Fellowship.



Israel Cohen (M'01–SM'03–F'15) is a Professor of electrical engineering at the Technion - Israel Institute of Technology, Haifa, Israel. He received the B.Sc. (Summa Cum Laude), M.Sc. and Ph.D. degrees in electrical engineering from the Technion - Israel Institute of Technology, in 1990, 1993 and 1998, respectively.

From 1990 to 1998, he was a Research Scientist with RAFAEL Research Laboratories, Haifa, Israel Ministry of Defense. From 1998 to 2001, he was a Postdoctoral Research Associate with the Computer Science Department, Yale University, New Haven, CT, USA. In 2001, he joined the Electrical Engineering Department of the Technion. He is a coeditor of the Multichannel Speech Processing Section of the *Springer Handbook of Speech Processing* (Springer, 2008), a coauthor of *Noise Reduction in Speech Processing* (Springer, 2009), a Coeditor of *Speech Processing in Modern Communication: Challenges and Perspectives* (Springer, 2010), and a General Cochair of the 2010 International Workshop on Acoustic Echo and Noise Control (IWAENC). He served as Guest Editor of the *European Association for Signal Processing Journal on Advances in Signal Processing* Special Issue on Advances in Multimicrophone Speech Processing and the *Elsevier Speech Communication Journal* a Special Issue on Speech Enhancement. His research interests are statistical signal processing, analysis and modeling of acoustic signals, speech enhancement, noise estimation, microphone arrays, source localization, blind source separation, system identification and adaptive filtering.

Dr. Cohen was a recipient of the Alexander Goldberg Prize for Excellence in Research, and the Muriel and David Jacknow Award for Excellence in Teaching. He serves as a member of the IEEE Audio and Acoustic Signal Processing Technical Committee and the IEEE Speech and Language Processing Technical Committee. He served as Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and IEEE SIGNAL PROCESSING LETTERS.