Multimodal Kernel Method for Activity Detection of Sound Sources

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

***Abstract*—We consider the problem of acoustic scene analysis of multiple sound sources. In our setting, the sound sources are measured by a single microphone, and a particular source of interest is also captured by a video camera during a short time interval. The goal in this paper is to detect the activity of the source of interest even when the video data are missing, while ignoring the other sound sources. To address this problem, we propose a kernel-based algorithm that incorporates the audio–visual data by a combination of affinity kernels, constructed separately from the audio and the video data. We introduce a distance measure between data points that is associated with the source of interest, while reducing the effect of the other (interfering) sources. Using this distance, we devise a measure for the presence of the source of interest, which is naturally extended to time intervals, in which only the audio signal is available. Experimental results demonstrate the improved performance of the proposed algorithm compared to competing approaches implying the significance of the video signal in the analysis of complex acoustic scenes.**

***Index Terms*—Acoustic scene, audio-visual, data fusion, kernel, multi-modal, transient noise.**

I. Introduction

A KEY element of automatic systems analyzing sound scenes is the ability to distinguish between different sound sources, which are often active simultaneously. In this paper, we consider sound sources of different types including speech, stationary and quasi-stationary background noises, as well as transient interferences, which are abrupt sounds, such as door-knocks and keyboard taps [1]. The sound sources are measured by a single microphone. In addition, a particular sound source is measured by a video camera, which is used as a "spotlight" to designate the source of interest. Examples of video frames of sources of interest are presented in Fig. 1, and they include speech, keyboard tapping and drum beats. The objective in this work is to detect the time intervals in which the source of interest is active. We consider a challenging setting, where the audio-visual recording is available only for a short time period, while in the remainder of the time, only the audio signal, which is processed in an online manner, is available. In addition, the detection is performed in an unsupervised manner, such that we do not have the true labels of the sources.

Detecting the activity of a source of interest may be very useful for sound scene analysis. For example, the scene may be decomposed into its components in a step by step procedure. At each step, the video camera is pointed at a particular source, enabling to learn to identify the activity of this particular source from the complex audio recordings. Pointing the video camera to a certain source of interest may be seen as an "automatic focusing" procedure, which is analogous to the human audio perception guided by visual inputs. Considering the availability of the video data only in a limited time interval is particularly practical for simultaneous activity detection of multiple sound sources. Since, by assumption, the video camera can measure merely a single sound source at a time, one may gradually and separately collect video data from each sound source, and, as we show, use the recording of a particular source for improving its activity detection even when the video data are no longer available.

The activity detection of sources of interest may be further useful for applications such as speech enhancement. Consider for example the enhancement of speech measured by a single microphone and a web camera during a voice over IP (VOIP) conversation in the presence of keyboard taps. A common key procedure in speech enhancement systems is the accurate detection of the presence of speech and the interferences [2], [3], which is carried out in this paper by the incorporation of the video camera. Since collecting the video of speech and the keyboard taps simultaneously is not practical using a single video camera, the data of these sources are collected one by one during a short "calibration" time interval, and in testing time intervals the data (of at least one of them) is missing. Moreover, assuming that the video data are only partially available, it is beneficial in real life scenarios such as sudden degradation of the video signal. For example, the speaker may move his head out of the video frame during natural speech.

Related problems dealing with the analysis of sound scenes are audio and audio-visual scene classification and event detection. Given an audio or audio-visual event, the goal is to assign it with the most appropriate class selected from a finite set of classes, where a class of studies assume a monophonic setting in which only a single audio event is present in each time interval [4]–[10]. The present work belongs to a recent line of studies dealing with a polyphonic setting, where multiple sounds may be active simultaneously [11]–[16]. There are several

Manuscript received July 11, 2016; revised November 5, 2016 and January 16, 2017; accepted January 25, 2017. Date of current version May 23, 2017. This work was supported by the Israel Science Foundation under Grants 576/16 and 1490/16). The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Gael Richard. *(Corresponding author: David Dov.)*

The authors are with the Andrew and Erna Viterby Faculty of Electrical Engineering, 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.2017.2690568

Fig. 1. Examples of video frames of sources of interest. From left to right: speech, drum beats, keyboard-tapping.

significant differences between these studies and the problem we consider here. First, in event detection, the types of sounds, i.e., the classes, are assumed to be known in advance. Second, in contrast to the current work where we use only the recorded unmarked data, large labeled databases are typically required to train the classifiers. For example, the authors in [17] reported that sound event classifiers based on deep neural networks could not outperform a baseline system based on a Gaussian mixture model on the DCASE dataset [18], due to the lack of sufficient amount of training data. Last, the annotation of the datasets requires significant human effort especially in the polyphonic case, since each time segment is annotated with multiple labels according to the multiple sound classes.

The methodology we present is based on obtaining a representation of the audio-visual signal in which the effect of the interfering sources is reduced. Related studies which are also based on unsupervised learning of representations of audio-visual signals were presented in [19]–[24]. In [20], the authors proposed to use mutual information as a measure of synchronization between audio and video features assuming the distribution of the signals follows a Gaussian model. Mutual information was also exploited in [19], where the authors suggested to map audio and video signals into domains designed to maximize the mutual information between the modalities. The authors in [21] proposed to obtain a representation of the audio-visual signal via a variant of the well-known Canonical Correlation Analysis (CCA) relying on the sparsity of events occurring simultaneously in both modalities. The methods presented in [23], [24] rely on the incorporation of the audio and the video signals via a simultaneous factorization of two non-negative matrices – one for each modality, applying the method to the problem of speaker diarization. Although the representation in these studies [19]–[24] is obtained in an unsupervised manner, they have two main limitations in the setting we consider. First, these representations are mainly learned via time-consuming solutions of optimization problems. Therefore, they are less suitable for obtaining a representation from a short sequence. Second, in contrast to this work, they assume that both the audio and the video modalities are available during the entire time.

We address the problem of the activity detection of the source of interest from a kernel-based geometric standpoint, in which the goal is to obtain a representation of the audio-visual data that respects relations between data points only in terms of the source of interest. Typical kernel-based geometric methods are designed for non-linear dimensionality reduction of single-modal data [25]–[29]. They provide low dimensional representations by the eigenvalue decomposition of affinity kernels aggregating local relations (affinities) between data points. Recent extensions of kernel methods to the multi-modal settings suggest constructing separate affinity kernels for each modality (audio and video in our case), and fusing the modalities through different combinations of the affinity kernels [30]–[43].

A particular data fusion approach, which is based on combining the data via the product of affinity kernels, was recently studied in [41]–[43]. In [42], we analyzed this fusion scheme in a discrete setting using graph theory. We viewed the single-modal affinity kernels and the product of kernels as defining single and multi-modal graphs, respectively, and studied the appropriate selection of their bandwidth, which are directly related to the graph connectivity and have a significant influence on the overall performance. In [41], Lederman and Talmon analyzed this fusion approach in a continuous setting, in which the affinity kernels are viewed as two diffusion operators, which are applied in an alternating manner. They showed that modality-specific factors, i.e., factors which appear only in one of the modalities, are attenuated by the alternation of steps.

In this paper, we propose an algorithm for the activity detection of sources of interest based on combining partially available audio and video signals, recorded over a short time interval. The algorithm exploits short synchronized sequences of audio and video signals incorporating the two modalities based on the method presented in [41], [42], where they are combined via the product of affinity kernels, constructed separately for each modality. The incorporation of the video signal improves the discriminative power of the unified affinity kernel, and it allows to construct a data-driven distance based on the unified kernel. This distance preserves relations between data points according to the source of interest, and it reduces the effect of other sound sources, which are modality (in our case, audio) specific. Using this distance, we devise a measure for the presence of the source of interest, which serves as a proxy for source activation labels in the absence of actual labels. Then, we show how to extend this measure to frames in which only the audio signal is available while preserving the properties of the data-driven distance. We apply the proposed algorithm to the detection of different types of sound sources including speech, drum beats and keyboard tapping, and examine its performance in challenging scenarios, in which the interferences are of a similar type as

the source of interest. The proposed algorithm attains improved performance compared to competing single- and multi-modal approaches demonstrating a significant contribution of the fusion of partially available audio-visual signals for sound scene analysis.

The contributions of this paper with respect to our previous work presented in [42] is as follows. First, we address here the fusion problem of *partially available* audio-visual signals in an *online setting* in contrast to the batch setting, with fully available signals, which was considered in [42]. As far as we know, this paper is the first to demonstrate a successful extension of the fusion method presented in [41], [42] to partially available multi-modal signals, i.e., signals measured by sensors of different types (audio and video). In addition, in [42], we have focused on the graph theoretic analysis of this fusion approach, and only demonstrated it for the problem of voice activity detection, which is a relatively simple special case of the problem we consider here. The much wider task of sound source activity detection, considered in this paper, includes not only different types of sources and multiple simultaneous interferences, but also cases where the source of interest and the interferences are of the same type, e.g., both are speech from different speakers or taps from different keyboards. Specifically, the activity detection of other sources rather than speech, e.g., keyboard taps, was not addressed in the literature, to the best of our knowledge. We further note that the analysis of the video signal of the different types of sources may be considered as different tasks from a computer vision point of view. For the analysis of speech signals, for example, complex algorithms are often used to accurately detect and track key-points in the mouth region of the speaker [44]–[46], and they cannot be directly applied for the detection of keyboard taps. Moreover, as we show, constructing a measure of activity based merely on the video signal leads to poor detection results especially in the detection of sources other than speech. Yet, the different video signals are handled in a similar manner by our proposed algorithm for the detection of the presence of a broad variety of sources of interest.

The remainder of the paper is organized as follows. In Section II, we formulate the problem. In Section III, we propose an algorithm for activity detection of sources of interest, and present experimental results demonstrating its improved performance in Section IV.

II. Problem Formulation

Consider a complex acoustic scene comprising multiple sound sources, such as speech, different types of transients and background noises, which may be active simultaneously. The acoustic scene is measured by a single microphone, and the measured signal is processed in frames. Let $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_N$ be a feature representation of a sequence of N frames, where $\mathbf{a}_n \in \mathbb{R}^{P_a}$ is the nth time frame, and P_a is the number of features, which are described in Section IV. Assuming $R+1$ audio sources, denoted by $s_1, s_2, ..., s_R, \tilde{s}$, the audio signal is viewed as an unknown (possibly) non-linear mapping f of the sources:

$$\mathbf{a}_n = f(s_1^a, s_2^a, ..., s_R^a, \tilde{s}).$$

The acoustic scene is also captured by a video camera, which is used as a "spotlight" that designates the source $\tilde{s}$ whose presence we would like to detect. We term the source $\tilde{s}$ "the source of interest" and consider all other R sources as interferences. Let $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_L$ be a sequence of L video frames, where $\mathbf{v}_n \in \mathbb{R}^{P_v}$ is a features representation of the nth frame. We consider a setting, in which the video signal is available only in a subset of the time interval of the audio signal, i.e., $L < N$. The sequence of the video frames is aligned to the audio sequence $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_L$ by a proper selection of the frame length and the overlap of the audio signal as described in Section IV. The video signal may also contain interfering sources, so that the video signal is seen as an unknown mapping g of the sources:

$$\mathbf{v}_n = g(s_1^v, s_2^v, ..., s_Q^v, \tilde{s}),$$

where we assume Q interfering source, $s_1^v, s_2^v, ..., s_Q^v$ [1]. For example, when the camera is pointed at the face of a speaker, head movements are considered interferences since they are not directly related to the production of speech. The only source measured by both the video camera and the microphone is the source of interest such that all other sources are assumed modality specific, an assumption that we use in Section III to construct a measure of the presence of the source of interest.

Let $\mathcal{H}_0$, $\mathcal{H}_1$ be hypotheses of the absence and the presence of the source of interest $\tilde{s}$, respectively, and let $\mathbb{1}_n$ be the corresponding indicator of the nth frame, given by:

$$\mathbb{1}_n = \left\{ \begin{array}{l} 1,\ n \in \mathcal{H}_1 \\ 0,\ n \in \mathcal{H}_0 \end{array} \right\}. \tag{1}$$

The goal in this paper is to detect the presence of the source of interest, while ignoring all other sources, i.e., to estimate $\mathbb{1}_n$ in (1). Specifically, we focus on estimating the indicator $\mathbb{1}_n$ in time intervals, in which the video signal is missing, i.e., $n \in [L+1, L+2, ..., N]$, and consider an online setting, where these frames are processed sequentially. We note that we consider an entirely unsupervised process of the estimation of $\mathbb{1}_n$ in (1) such that even for the interval $1, 2, ..., L$ we do not have labels indicating the presence of the sources.

III. Kernel-Based Detection of the Source of Interest

A. Audio-Visual Fusion via a Product of Affinity Kernels

We exploit the audio-visual data to construct a measure of the presence of the source of interest by fusing the data via a product of affinity kernels constructed separately for each modality. Let $\mathbf{K}^a \in \mathbb{R}^{L \times L}$ be an affinity kernel constructed from the sequence of audio frames $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_L$ such that its (n,m)th entry is given by:

$$K_{n,m}^a = \exp\left[- \left\| \mathbf{a}_n - \mathbf{a}_m \right\|_2^2 / \epsilon^a \right], \tag{2}$$

where $\|\cdot\|_2$ is the L_2 distance, and ϵ^a is the kernel bandwidth, a parameter whose selection we studied in [42]. The affinity kernel has an interpretation of a graph on the data, which we term the audio graph, whose nodes are the data points $\{\mathbf{a}_n\}$,

[1] Throughout this paper, a and v denote audio and video, respectively.

and the weight of the edge between node n and node m is given by $K^a_{n,m}$. Let $\mathbf{D}^a \in \mathbb{R}^{L\times L}$ be a diagonal matrix, whose nth element on the diagonal, denoted by $D^a_{n,n}$, is given by:

$$D^a_{n,n} = \sum_{m=1}^{L} K^a_{n,m}. \tag{3}$$

The matrix $\mathbf{D}^a$ is often referred to as the degree matrix, when the affinity function $K^a_{n,m}$ consists of binary values, so that $D^a_{n,n}$ is the number of vertices connected to vertex n. Here, we use the inverse of $\mathbf{D}^a$ to normalize the rows of $\mathbf{K}^a$ constructing a row stochastic matrix $\mathbf{M}^a \in \mathbb{R}^{L\times L}$ by:

$$\mathbf{M}^a = \left(\mathbf{D}^a\right)^{-1}\mathbf{K}^a. \tag{4}$$

The row stochastic matrix $\mathbf{M}^a$ defines a Markov chain on the graph such that its (n,m)th entry, denoted by $M^a_{n,m}$, represents the probability of transition from node n to node m in a single step. These transition probabilities incorporate information on the inter-relations between the samples/nodes. For example, in many manifold learning and kernel-based techniques, such as [29], they are used, via the eigenvalue decomposition, to obtain a global representation of the data.

The data from the two modalities are combined by the construction of the matrix $\mathbf{M} \in \mathbb{R}^{L\times L}$, which incorporates the data from the two modalities via the product of kernels:

$$\mathbf{M} = \mathbf{M}^a\mathbf{M}^v, \tag{5}$$

where $\mathbf{M}^v \in \mathbb{R}^{L\times L}$ is a row stochastic matrix constructed from the video signal, similarly to $\mathbf{M}^a$ according to (2)-(3). The matrix $\mathbf{M}$ is also row stochastic, so it defines an audio-visual graph, whose nodes correspond to the pairs of frames $(\mathbf{a}_1,\mathbf{v}_1),(\mathbf{a}_2,\mathbf{v}_2),...,(\mathbf{a}_L,\mathbf{v}_L)$. According to (5), the (n,m)th entry of $\mathbf{M}$ is explicitly given by:

$$M_{n,m} = \sum_{l=1}^{L} M^a_{n,l}M^v_{l,m}.$$

Therefore, it may be interpreted as the probability of transitioning from node n to node m in two steps: first from node n to node l in the audio graph and then from node l to node m in the video graph. In the same sense, Lederman and Talmon showed in [41] that the continuous counterpart of $\mathbf{M}$ is a diffusion operator employing two diffusion steps, one for each modality. They showed that such alternating diffusion steps attenuate the view specific factors, which are defined as interferences in our case. In SubSection III-B, we provide more insight on this result by describing the relation between the product of kernels and the diffusion distance [29], which in turn motivates us to build a measure for the presence of the source of interest as we describe in SubSection III-C.

B. Diffusion Distance

Let $d(n,m)$ be the diffusion distance between frame n and frame m, given by [41]:

$$d(n,m) = \sqrt{\sum_{l=1}^{N}\left(M_{n,l} - M_{m,l}\right)^2}. \tag{6}$$

According to (6), the distance between frame n and frame m is roughly given by a collection of transition probabilities in one step between the frames. Note that $d(n,m)$ is an unnormalized spacial case of the more general diffusion distance, presented in [29], comprising transition probabilities between frames in multiple steps. Since the distance between a pair of frames takes into account other frames in the set, the diffusion distance respects the geometry of the data and is considered robust to noise [29]. In addition, in the multi-modal setting we consider here, the diffusion distance is constructed from the matrix $\mathbf{M}$, so that it measures distances between frames according to both the audio and the video sources, $s^a_1, s^a_2, ..., s^a_R, s^v_1, s^v_2, ..., s^v_Q, \tilde{s}$.

The diffusion distance may be rewritten in terms of a distance between two vectors corresponding to frame n and frame m. Specifically, let $\mathbf{h}_n \in \mathbb{R}^L$ be a vector corresponding to frame n, given by:

$$\mathbf{h}_n = \mathbf{M}^T\mathbf{h}^0_n,$$

where T denotes transpose, and $\mathbf{h}^0_n \in \mathbb{R}^L$ is an indicator vector whose nth element equals one and all other elements equal zero. Accordingly, the diffusion distance $d(n,m)$ in (6) is given by:

$$d(n,m) = \left\|\mathbf{h}_n - \mathbf{h}_m\right\|_2. \tag{7}$$

The use of the product of kernels for the fusion of the audio and the video signals is motivated by [41, Th. 5], presented in the continuous domain, implying on the existence of equivalent functions to $\mathbf{h}_n$ and $\mathbf{h}_m$, which are merely functions of the source of interest $\tilde{s}$. Namely, on the one hand, the diffusion distance is a data driven distance that can be explicitly calculated for each pair of frames according to (6). On the other hand, it is equivalent to a distance between implicit functions, which are functions of merely the source of interest, so that it allows measuring distances between data points in terms of the source of interest only, while ignoring all other sources, which are modality-specific by assumption. For more details, we refer the readers to [41].

C. Detection of the Presence of the Source of Interest

The proposed measure of the presence of sources of interest is constructed from the eigenvalue decomposition of the matrix $\mathbf{M}$ in (5). Since the matrix $\mathbf{M}$ is row stochastic, it has an all ones eigenvector corresponding to the eigenvalue 1, which is ignored since it does not contain information. Let $\boldsymbol{\phi}_1,\boldsymbol{\phi}_2,...,\boldsymbol{\phi}_{L-1}$ and $\lambda_1,\lambda_2,...,\lambda_{L-1}$ be the eigenvectors (excluding the trivial) and the corresponding eigenvalues of $\mathbf{M}$, respectively. The motivation to use the eigenvalue decomposition of $\mathbf{M}$ for the detection of the presence of the source of interest stems directly from its relation to the diffusion distance [29], [41]:

$$d(n,m) = \sqrt{\sum_{l=1}^{N}\lambda_l\left(\phi_l(n) - \phi_l(m)\right)^2}, \tag{8}$$

where $\phi_l(n)$ is the nth entry of $\boldsymbol{\phi}_l$. The expression in (8) implies that the eigenvectors of the kernel product $\mathbf{M}$ may be used as new coordinates of the data samples representing them in terms of the source of interest. Since in this study we are only

interested in the estimation of a single indicator, we use only the leading eigenvector $\boldsymbol{\phi}_1$. Specifically, we propose to estimate the indicator of the source of interest in frame $n \in [1, 2, ..., L]$, $\mathbb{1}_n$ in (1), by:

$$\hat{\mathbb{1}}_n = \begin{cases} 1 & ; \quad \phi_1(n) > \tau \\ 0 & ; \quad \text{otherwise} \end{cases}, \tag{9}$$

where τ is a threshold value. We note that the leading eigenvector is of length L as the number of the frames from which it is constructed, such that its nth entry corresponds to the nth data point. The leading eigenvector of a row stochastic matrix is often used in the literature for clustering since it solves the well-known normalized cut problem; specifically, the nth data point is assigned to one of two possible clusters according to the sign of the corresponding nth entry of the leading eigenvector [47]. In our case, the leading eigenvector of the unified affinity kernel $\mathbf{M}$ clusters the signal according to the presence of the source of interest, and indeed, as we show in Section IV, high values of the entries of this eigenvector correspond to frames, in which the source of interest is active, while low values are obtained for inactive frames. In addition, we use the leading eigenvector as a continuous measure, such that thresholding allows us to control the trade-off between correct detection and false alarm rates. For example, low threshold values should be set in applications where high detection rates are required at the expense of higher rates of false alarms; when no addition information is available on the signal or the application at hand, the threshold may be set to zero to cluster the signal according to the sign of the entries as proposed in [47].

Two additional properties make the leading eigenvector $\boldsymbol{\phi}_1$ particularly useful for the detection of sources of interest; first, it is constructed in a data-driven manner, so that the indicator of the presence of the source of interest, $\hat{\mathbb{1}}_n$ in (9), is estimated without any other information. Specifically, the true labels of the presence of the source of interest are not required.

Second, the eigenvector may be extended to frames $L+1$, $L+2, ..., N$ even though they comprise only audio data [43], [48], as we describe next. Given a new frame $\mathbf{a}_n$, $n \in [L+1, L+2, ..., N]$, we use the nyström method [49] to obtain a new entry of $\boldsymbol{\phi}_1$ corresponding to frame n, which is denoted by $\phi_1(n)$:

$$\phi_1(n) = \frac{1}{\lambda_1} \sum_{m=1}^{L} M_{n,m} \phi_1(m). \tag{10}$$

By (5), (10) can be rewritten as:

$$\phi_1(n) = \frac{1}{\lambda_1} \sum_{m=1}^{L} \sum_{l=1}^{L} M^a_{n,l} M^v_{l,m} \phi_1(m) \triangleq \sum_{l=1}^{L} M^a_{n,l} \theta(l), \tag{11}$$

where $\theta(l) = \frac{1}{\lambda_1} \sum_{m=1}^{L} M^v_{l,m} \phi_1(m)$. The right term in (11) implies that given a new frame n, the extension requires only the audio frame $\mathbf{a}_n$ since the term $\theta(l)$, which comprises the video (and the audio) data, is calculated based only on frames $1, 2, ..., L$.

At this point, we note that the matrices $\mathbf{M}^a$ and $\mathbf{M}^v$ are similar to symmetric matrices, so that their eigenvectors are guaranteed to be real-valued [29], which is not the case for $\mathbf{M}$. One solution for this problem is to use the singular value decomposition of $\mathbf{M}$, which is shown by Lindenbaum *et al.* in [40] to provide another variant of the diffusion distance. Yet, we use in this study the leading eigenvector instead of, e.g., the leading singular vector, since (i) the leading eigenvector indeed appears real-valued in all our experiments, (ii) it may be extended to new incoming frames using the nyström method, and (iii) it provides better detection results. We summarize the proposed algorithm for the detection of the presence of the source of interest in Algorithm 1.

Algorithm 1: Detection of the presence of the source of interest

Obtain the first L pairs of frames $\{\mathbf{a}_n, \mathbf{v}_n\}_{n=1}^{L}$
Calculate the affinity kernels $\mathbf{K}^a$ and $\mathbf{K}^v$ according to (2)
Calculate the row stochastic matrices $\mathbf{M}^a$ and $\mathbf{M}^v$ according to (3)–(4)
Fuse the data via the product of kernels, i.e., compute $\mathbf{M}$ according to (5)
Obtain the leading eigenvector $\boldsymbol{\phi}_1$
Extension to frames $L+1, L+2, ...$
for $n = L+1, L+2, ...$ **do**
    Obtain the audio frame $\mathbf{a}_n$
    Calculate affinities to frames $1, 2, ..., L$: $\left\{M^a_{n,l}\right\}_{l=1}^{L}$
    Calculate the new entry of the eigenvector $\phi_1(n)$ using (11)
    if $\phi_1(n) > \tau$ **then**
        $\hat{\mathbb{1}}_n = 1$
    else
        $\hat{\mathbb{1}}_n = 0$
    end if
end for

IV. Experimental Results

A. Experimental Setting

To evaluate the performance of the proposed algorithm we use audio and audio-visual recordings[2] of different types of sound sources including speech, different types of noise and transients, which are synthetically added (in the audio modality) to simulate complex audio scenes with multiple sources. Each recording is divided into two parts of equal lengths such that the first part comprises both the audio and the video, and the second part comprises only the audio. The second part of the recordings with the missing video data is processed in an online manner and is used for the evaluation of the algorithm.

Each recording is a sequence of 90–120 s length, sampled by the video camera at 25–30 fps. The audio signal is sampled at 8 kHz and processed in frames with 50% overlap, where the frame length is set to $\sim$630 samples such that the audio frames are aligned with the video frames. To evaluate the performance

[2]The audio and audio-visual recordings are available at https://davidov312.github.io/ADMrefSet/

of the proposed method, we use the clean audio recording of the source of interest. We set the ground truth for the true presence of the source of interest by comparing the energy of the clean signal to a threshold whose value is set to 1% of the maximal energy value in the sequence. The source of interest is considered present in frames with energy value above this threshold value. In this challenging type of ground truth setting, transitions between the presence and the absence of the source of interest may occur in the resolution of tens of ms.

For the representation of the audio signal, we use the Mel-Frequency Cepstral Coefficients (MFCC), which are calculated by filtering the audio signal in the domain of the power spectra with a bank of the perceptually meaningful Mel-scale filters. The MFCC representation is given by the coefficients of the discrete cosine transform (DCT) applied to the log of the outputs of the filters. The MFCCs represent the spectrum of the signal in a compact form, and they are widely used in a variety of audio processing applications [50]–[52]. We use a Matlab implementation of the MFCCs, taken from [53], and set the number of coefficients to 24. We found in our experiments that the performance of our method is not sensitive to the particular number of coefficients. In addition, we set the number of filters to 90. We empirically found that the optimal number of filters depends on the type of the source of interest. When the source of interest has a more abrupt nature, e.g., keyboard taps, a larger number of filters should be used, and for more "stationary" signals, such as speech, a lower number of filters provide better performance. Since we do not assume in this study that the type of the source of interest is known, we use 90 filters, which is an intermediate value providing good performance for all types of sources of interest. In this context, we note that using a higher sampling rate than 8 kHz has a negligible effect on the performance.

In addition, we note that the effect of the feature selection process on the accuracy of the activity detection implies that their proper selection may lead to further improvement of the proposed algorithm. One approach, which we leave to a future study, would be to learn the features from the data, e.g., using deep learning methods based on unsupervised learning procedures such as deep belief networks [54]. However, such procedures should be applied offline, and since the type of the sources and interferences are not known in advance, a large database of sounds should be exploited.

The video signals have resolutions in the range of 328×184 to 640×480 pixels, and they are represented by motion vectors. We use a Matlab implementation of Lucas - Kanade method [55], [56] (vision.OpticalFlow Matlab system object) to estimate the motion of non-overlapping blocks of 10×10 pixels between pairs of consecutive frames. Then, we concatenate the absolute values of the motion in each block into vectors. The feature representation of frame n, $(\mathbf{a}_n, \mathbf{v}_n)$, is given by the concatenation of the motion vectors and the MFCCs in frames $n-1, n$ and $n+1$, respectively. The use of data from three consecutive frames for the representation of the audio-visual signal allows for the incorporation of temporal information into the proposed algorithm, which is not taken into account in the construction of the affinity kernels $\mathbf{M}^a$ and $\mathbf{M}^v$.

Before turning to the experimental results, we note that rather than extending the eigenvector $\boldsymbol{\phi}_1$ to a frame l, for which the video data is missing according to (11), a more computationally efficient extension is obtained by:

$$\phi_1(l) = \sum_{m=1}^{L} M^a_{l,m} \phi_1(m), \tag{12}$$

The extension in (12) may be seen as a weighted interpolation of the measure of the presence of the source of interest based only on the audio signal, which is the one available for new incoming frames. Specifically, since $\mathbf{M}^a$ is a row stochastic matrix, the "weights" $M^a_{l,m}$ sum to one, and the more similar frame $\mathbf{a}_l$ to a certain frame $\mathbf{a}_m$, $m \in 1, 2, ..., L$, the higher the corresponding weight $M^a_{l,m}$ is. In addition, we found in our experiments that the extension in (12) provides better results, so it is the one used in the reported results. In this context, we note that the eigenvalue decomposition assigns an arbitrary sign to the eigenvectors. We assume that the correct sign of the eigenvector $\boldsymbol{\phi}_1$ is known, and that high entry values correspond to frames in which the source of interest is present; in practice, the sign may be chosen such that a negative sign is assigned to entries of the eigenvector corresponding to frames, in which all audio sources are absent, i.e., silent frames.

Since the proposed approach is evaluated for frames in which the video data is missing, we compare it to an approach, which is based only on the audio data, in order to highlight the contribution of the video signal. Specifically, we compare the proposed method to its single modal variant, in which only the audio signal is exploited in frames $1, 2, ..., L$ for the construction of the measure of the presence of the source of interest; namely, the leading eigenvector of the matrix $\mathbf{M}^a$ is used to construct the measure. The single modal approach may be seen as an unsupervised variant of the method presented in [57], which is based on using eigenvectors of an affinity kernel for speech detection.

In addition, we compare the proposed algorithm to the Canonical Correlation Analysis (CCA) method, which is denoted by "CCA" in the plots, and to the method presented in [19]. The methods are based on obtaining representations of the the audio and the video signals by mapping them to new domains, in which the correlation and the mutual information between the modalities is maximized, respectively. The method presented in [19] is denoted in the plots by MMI (maximization of mutual information).

We also present the performance of a variant of the proposed algorithm based only on the video signal. This approach cannot be used in practice in the setting we consider since it requires the availability of the video signal in the evaluated time intervals, in which it is assumed missing. Still, the performance of the approach based only on the video data is presented to further gain insight into the contribution of the fusion procedure between the audio and the video data for the activity detection of the source of interest.

B. Activity Detection of Speech Sources

In the first experiment, we consider speech as the source of interest. We use an audio-visual dataset, which we

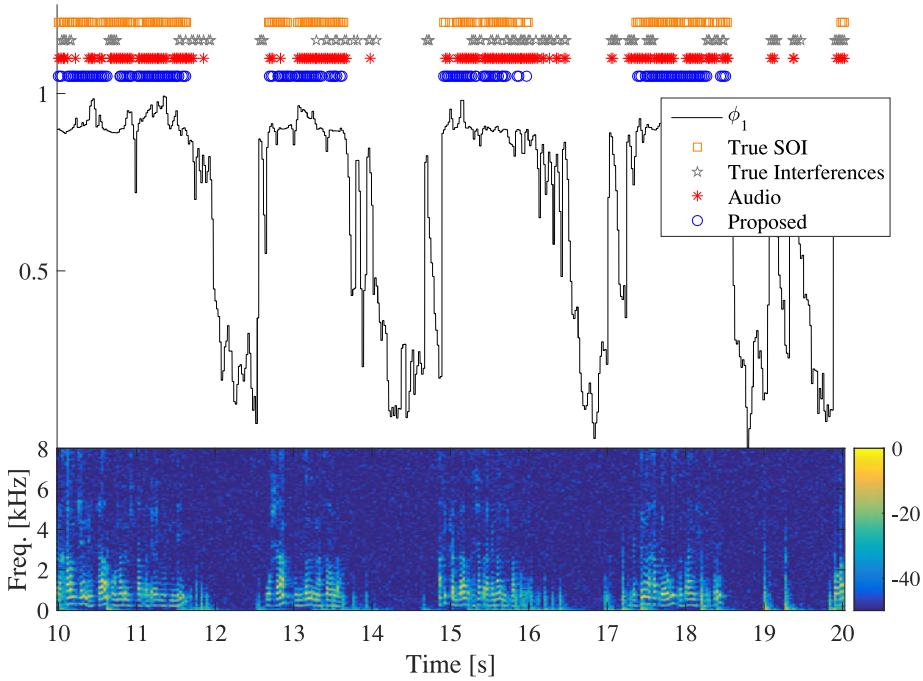


Fig. 2. Qualitative assessment of the proposed algorithm for the activity detection of the source of interest. Source of interest: speech. Interfering source: door-knock transients with SIR 1. (Top) Time domain, trajectory of the leading eigenvector - black solid line, true SOI (speech) - orange squares, true interferences (transients) - gray stars, a variant of the proposed method based only on the audio signal with a threshold set for 80% correct detection rate - red asterisks, proposed algorithm with a threshold set for 80% correct detection rate- blue circles. (Bottom) Spectrogram of the input signal.

presented in [58] comprising 11 sequences of different speakers recorded via a smartphone. We synthetically add different types of noise and transients taken from a free online corpus [59], with different SNRs and with different source of interest to interferences ratios (SIR). Specifically, we define the SIR as the ratio between the maximal amplitudes of the source of interest and the interferences (transients in this case) such that the SIR equals one when they have the same maximal amplitudes. We find this type of normalization based on the maximal amplitude more suitable than, e.g., using the power of the signals, due to the abrupt nature of the transients and it was previously used in [1]. The video signal comprises the face of the speaker, and it may comprise slight head and mouth movements in time intervals, in which speech, i.e., the source of interest, is absent.

An example of the detection of speech in the presence of door-knocks is presented in Fig. 2, where at the bottom of the figure we plot the spectrogram of the signal demonstrating the similar spectrum of the different audio sources, i.e., speech and the transients. In Fig. 2 at the top, we plot (black solid line) the proposed measure for the presence of the source of interest, $\phi_1(l)$, which is normalized to the range of $[0, 1]$ for the ease of presentation. Due to the normalization, it can also be viewed as the probability of the presence of the source of interest. It may be seen in Fig. 2 that the proposed measure properly provides high values in time intervals, in which the source of interest (speech) is indeed present. We compare the proposed approach with the audio-based approach to gain insight on the contribution of the video signal in the calibration set. We set the threshold value τ in (9) to provide 80% correct detection rate and compare their false alarms. It can be seen in Fig. 2 that the method based only on the audio signal provides more false alarms, e.g., around the 12th and the 17th sec.

We further evaluate the performance of the proposed method in Fig. 3 in the form of Receiver Operating Characteristic (ROC) curves, which are plots of the probability of detection versus the probability of false alarm. The curves are obtained by changing the threshold value τ in (9) over the value range of the measure of the presence of the source of interest ϕ_1. The higher the curve, i.e., the larger the Area Under the Curve (AUC), the better the performance of the corresponding method are. The AUC values are reported in the legend box for each method.

It may be seen in Fig. 3 that the proposed algorithm for the detection of sources of interest outperforms the competing methods. Specifically, the inferior performance of the variant based only on audio implies that using the video signal, the proposed algorithm indeed learns a measure of the presence of the source of interest, in which the effect of the interfering source is reduced, even though the video signal is missing in the evaluated time intervals. Therefore, the video signal allows for the analysis of the audio scene by properly distinguishing the sound source at which the video camera is pointed from all other sources.

The method based only on the video signal provides significantly inferior results to the proposed algorithm, which demonstrates that the video signal alone cannot provide accurate activity detection of the source of interest, even though it does not measure other sound sources in the scene. One reason for the inferior results is that the video signal may comprise visual cues which are not directly related to the source of interest, such

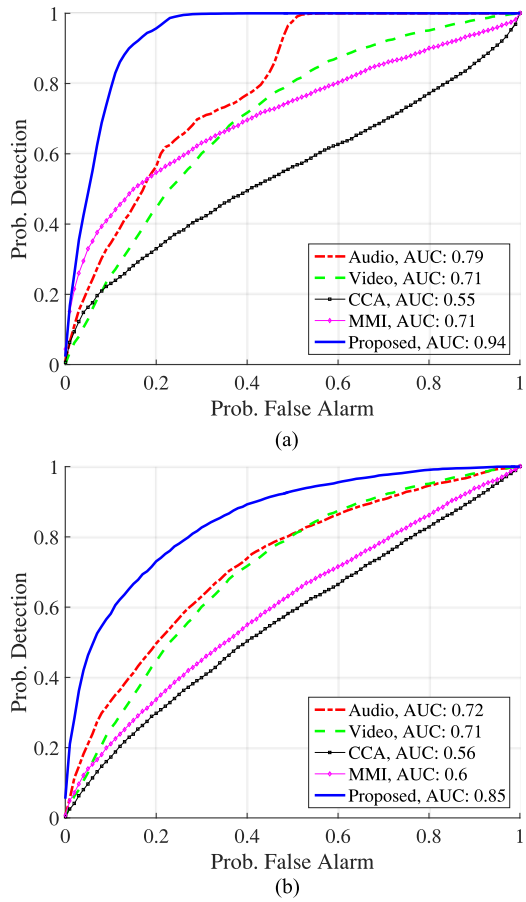


Fig. 3. Probability of detection vs probability of false alarm. Source of interest: speech. Interfering sources: (a) door-knock transients with SIR 1, (b) babble noise with 0 dB SNR and scissors transient with SIR 1.

as head movements of the speaker, which are seen as interfering sources.

In this context, we note that in a setting where both the audio and the video signals are available for a new incoming frame, the extension in (11) does not use the incoming video frame and its incorporation is an open problem, which we leave for a future study. Yet, we examine in our experiments a straightforward solution based on building the extension weights in (12) relying on similarities between unified audio-visual feature vectors constructed via the concatenation of the audio and the video features. Since we found that this alternative does not improve the detection scores, the corresponding results were discarded.

Moreover, we note that in [42], [60] we considered the fusion of audio-visual data using the product of kernels for speech detection. We showed that it provides better detection scores compared to alternative fusion schemes and the methods presented in [58], [61]. However, in [42], [60], we considered a batch setting, where the audio-visual data is available in advance; in contrast, here, we consider an online setting, in which only the audio data is available in the evaluated time intervals. In addition, in [42], we considered a cropped region of the mouth of the speaker as the video signal, assuming that accurate detection of the mouth region is required as a preprocessing stage. Instead, in this study we use the whole video recording including the face of the speaker, which pose a challenge since, e.g., movements of the head of the speaker may degrade the detection. Fig. 3 demonstrates that the proposed algorithm significantly outperform the alternative approaches.

We summarize the AUC scores of the different methods in the detection of speech in Table I (a-c) for different SIR levels. Table I comprises also the statistics of the activity of the different sources including the total number of the tested frames; the number of frames comprising the source of interest; the number of frames comprising the interferences; and those containing both of them. The statistics of the interfering sources account for the transients and speech but not for the stationary noise since the latter appears in all of the frames. We note that speech is a different type of sound compared to the interfering sources such as (quasi-) stationary babble noise or, e.g., the abrupt varying door-knocks. We further present in Table I the performance of the methods in the detection of speech in the presence of another (interfering) speech source. The challenge in the detection of the source of interest in such a scenario is emphasized by the degradation of the performance of all methods. Still, the proposed method provides improved performance compared to all other methods.

C. Activity Detection of Transient Sources

We proceed with the demonstration of the performance of the proposed algorithm for other acoustic scenes with different sources of interest. In Figs. 4 and 5, we use an audio-visual recording of drum beats and 7 audio-visual recordings of keyboard-taps, respectively, all taken from YouTube. The recordings of keyboard taps comprise different keyboards recorded from different angles. The corresponding audio sources are pre-filtered by the algorithm proposed in [2] to reduce stationary noise. As an interfering source in these experiments, we use, in addition to transients, speech signals taken from TIMIT database [62].

We note that the detection of the presence of these types of sources is significantly more challenging than speech activity detection. First, the sources of interest are present in very short time intervals of up to a single frame such that incorporating temporal information is not useful. Second, the audio scene comprises speech, which is a complex and a non-stationary interfering source spanning large ranges of amplitude and frequency values. Third, as far as we know, the detection of the presence of such sources is not studied in the literature in the setting we consider here, where the only available prior information is a short unmarked audio-visual recording. Last, the video signal of the different types of the sources, e.g., speech and keyboard taps, visually differs from each other as demonstrated in Fig. 1.

Fig. 4 demonstrates the accurate detection of drum beats in the presence of interfering speech. We consider the drum beats as an example of challenging audio-visual cues with complex relations between the audio and the video modalities. Specifically, the video features capture mainly the movement of the drumsticks; these cues are not equivalent to the production of sound, since sounds occur only in very short time intervals, when the sticks hit the drums, while the sticks move also before and

TABLE I
(A)–(C) AUC SCORES

Interfering sources	Audio	Video	CCA	MMI	Proposed
Door-knock transients	0.79	0.71	0.59	0.67	**0.94**
Babble noise with 0 dB SNR, scissors transient	0.73	0.71	0.56	0.57	**0.85**
Speech, babble noise with 20 dB SNR, door-knock transients	0.74	0.71	0.54	0.58	**0.79**

(a)

Interfering sources	Audio	Video	CCA	MMI	Proposed
Door-knock transients	0.91	0.71	0.53	0.85	**0.95**
Babble noise with 0 dB SNR, scissors transient	0.75	0.71	0.54	0.63	**0.87**
Speech, babble noise with 20 dB SNR, door-knock transients	0.79	0.71	0.54	0.61	**0.86**

(b)

Interfering sources	Audio	Video	CCA	MMI	Proposed
Door-knock transients	0.69	0.71	0.56	0.66	**0.89**
Babble noise with 0 dB SNR, scissors transient	0.67	0.71	0.53	0.58	**0.83**
Speech, babble noise with 20 dB SNR, door-knock transients	0.68	0.71	0.56	0.61	**0.73**

(c)

Interfering sources	Number of interfering frames	Number of frames containing both the source of interest and interferences
Door-knock transients	4778 (29%)	1578 (9%)
Babble noise with 0 dB SNR, scissors transient	8043 (43%)	2429 (15%)
Speech, babble noise with 20 dB SNR, door-knock transients	8781 (53%)	2891 (17%)

(d)

Source of interest: speech. SIR: (a) 1, (b) 2, (c) 0.5. Number of tested frames: 16665. Number of frames containing the source of interest: 5560 (33%). (d) Statistics on the activity of the interferences.

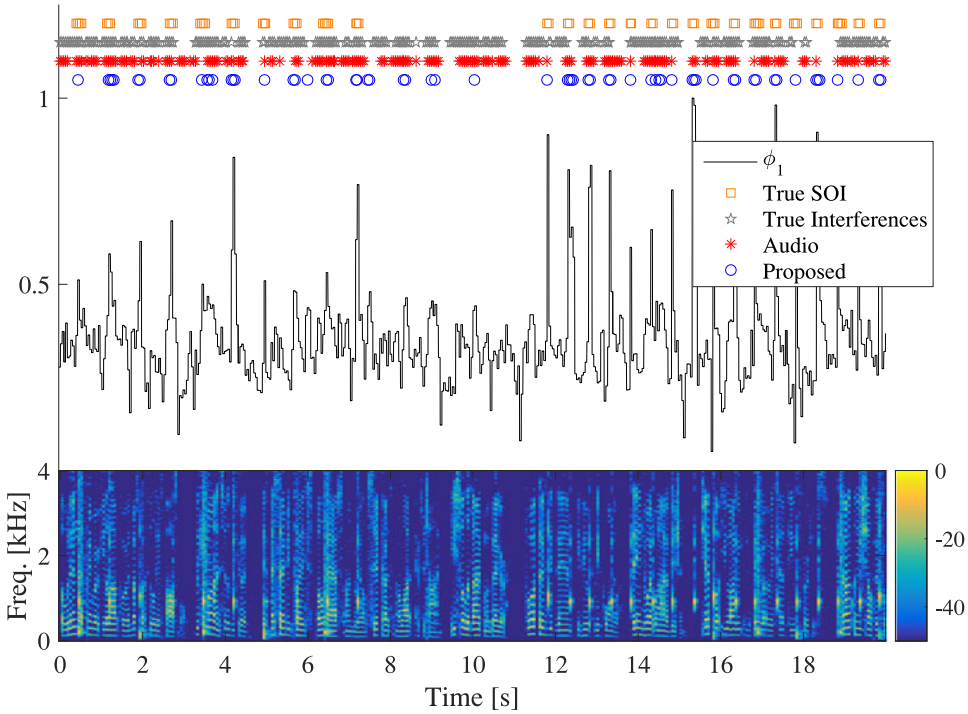


Fig. 4. Qualitative assessment of the proposed algorithm for the activity detection of the source of interest. Source of interest: drum beats. Interfering source: speech with SIR 2. (Top) Time domain, trajectory of the leading eigenvector - black solid line, true SOI (drum beats) - orange squares, true interferences (speech)- gray stars, a variant of the proposed method based only on the audio signal with a threshold set for 80% correct detection rate - red asterisks, proposed algorithm with a threshold set for 80% correct detection rate- blue circles. (Bottom) Spectrogram of the input signal.

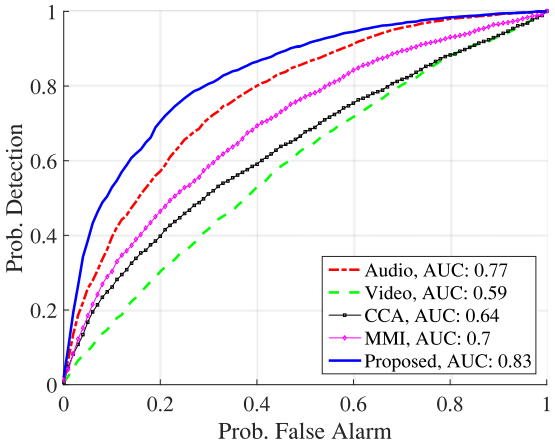


Fig. 5. Probability of detection vs probability of false alarm. Source of interest: keyboard taps. Interfering source: speech with SIR 2.

after these events. We observe that the proposed measure for the detection of the source of interest indeed provides high peaks in time frames, in which the drum beats indeed produce sound, since in these frames the source of interest is active simultaneously in both modalities. We further observe that the source of interest may be present for short time intervals, of single frames, a regime, which significantly differs from the speech as can be seen in Fig. 2. Yet, the proposed algorithm successfully detects these different sources of interest since it is mainly based on the affinities between the frames and not on a temporal information. Moreover, the proposed algorithm provides fewer false alarms compared to the method based only on the audio signal demonstrating the advantage of the incorporation of the video signal.

In Fig. 5, we demonstrate the performance of the detection of keyboard taps in the presence of interfering speech. The detection of keyboard-taps is especially challenging since first, there are rapid transitions between its presence and absence, and second, the corresponding video signal comprises almost nonstop movements of the hands of the user. Moreover, we use videos, in which keyboard taps are recorded from different angles and distances; and in few of them, there exist partial occlusions, e.g., when certain fingers or parts of the hand occlude the other parts. Indeed, the performance of the variant of the proposed algorithm based on the video signal completely fails in indicating the presence of the keyboard taps. Yet, in such a case, the proposed algorithm provides improved performance compared to the alternative approaches. Namely, despite the challenge in the analysis of keyboard-taps using the video signal, and despite its absence in the tested time intervals, the proposed algorithm successfully incorporates the video signal outperforming the alternative approaches.

In Table II we present the performance of the different methods for the activity detection of keyboard-tapping in the presence of interfering sources with different levels of SIRs. In addition to speech, we consider also transient interferences, which are similar sounds to the keyboard taps including hammering and taps from another keyboard. To demonstrate the effect of these interferences, we set the SIR level of speech to two and vary only the levels of the transient interferences. The improved performance of the proposed method demonstrates the contribution of the incorporation of the partially available video signal via the product of kernels for improving the analysis of complex sound scenes.

D. Discussion

The ability to obtain a representation of audio-visual signals according to factors that are common to the two modalities gives rise to extending the proposed approach to other applications directly related to the analysis of sound scenes. For example, the proposed approach may be applied for speaker diarization, i.e., to the task of "who spoke when", by using multiple video cameras, each pointed at a different speaker. In this case, the activity of each speaker is obtained by fusing the video signal from the camera pointed to him/her with the audio of the entire scene. In this context, we note that the fusion process based on the product between the affinity kernels detects, by design, the activity of all common sources among the two modalities, so that a single camera is not sufficient for polyphonic detection as is. To overcome this limitation, one may incorporate, e.g., a face detection algorithm to locate the speakers within the video, then isolate the region of the video frame containing a particular speaker, and fuse it with the audio signal for the activity detection of this speaker.

Moreover, the proposed approach may be extended to the task of source localization in videos, e.g., by analyzing the effect of removing regions from the video signal before the fusion process. Specifically, since the parts of the video signal, in which the source of interest is not present, are assumed to contain merely interferences, removing them should have a negligible effect on the source activity pattern in contrast to removing parts of the video that indeed contain the source of interest. In the presence of multiple sources of interest, as in the case of speaker diarization from a single video camera, one may learn the spatio-temporal patterns of the activity of the sources within the video assuming that the sources are active independently of each other and located in different regions of the video frame.

Finally, while we consider here an unsupervised setting, where the video signal is completely missing in the tested time intervals, we will consider in a future research a setting in which both the labels and the video signal are (at least partially) available. In this context, we point out the work presented in [63] addressing the analysis of multi-modal scenes using a matrix completion framework in a supervised setting with partially available labels. The framework is based on the incorporation of training and testing data along with the available labels into a matrix whose missing elements correspond to the (missing) testing labels. Then, the missing elements of the matrix are estimated via the solution of an optimization problem assuming a linear model for the generation of the labels from the data. The proposed approach may be further extended to a similar setting by the incorporation of the unified affinity kernel into a transductive learning framework presented in [64]. In the latter case, labels in the testing set are estimated by iteratively diffusing training labels with the testing set according to similarities

TABLE II
(A)–(C) AUC SCORES

Interfering sources	Audio	Video	CCA	MMI	Proposed
Speech	0.67	0.59	0.62	0.67	**0.78**
Speech, hammering	0.65	0.59	0.68	0.71	**0.76**
Speech, hammering, keyboard	0.65	0.59	0.64	0.62	**0.7**

(a)

Interfering sources	Audio	Video	CCA	MMI	Proposed
Speech	0.77	0.59	0.64	0.69	**0.83**
Speech, hammering	0.64	0.59	0.68	0.7	**0.8**
Speech, hammering, keyboard	0.65	0.59	0.68	0.75	**0.76**

(b)

Interfering sources	Audio	Video	CCA	MMI	Proposed
Speech	0.59	0.59	0.61	0.62	**0.71**
Speech, hammering	0.65	0.59	0.64	0.62	**0.7**
Speech, hammering, keyboard	0.64	0.59	0.61	0.63	**0.65**

(c)

Interfering sources	Number of interfering frames	Number of frames containing both the source of interest and interferences
Speech	7614 (77%)	3600 (36%)
Speech, hammering	7862 (79%)	3686 (37%)
Speech, hammering, keyboard	8388 (85%)	3929 (40%)

(d)

Source of interest: keyboard-tapping. SIR: (a) 1, (b) 2, (c) 0.5. Number of tested frames: 9906. Number of frames containing the source of interest 4686 (47%). (d) Statistics on the activity of the interferences.

(relations) between the training and the testing samples. The fusion of the audio and the video data via the product of the affinity kernels may allow for an improved diffusion of the labels while reducing the interfering factors in the different modalities.

V. CONCLUSION

We have addressed the analysis of an acoustic scene comprising multiple sound sources using a single microphone and a video camera, which is used as a spotlight pointed to a particular source of interest. The proposed algorithm utilizes the audio and the video data, which is available only in a short time interval, through a product of affinity kernels, separately constructed for each modality. The leading eigenvector of the product of kernels is used as a data-driven measure for the presence of the source of interest, and it is extended in an online manner to time intervals in which only the audio data is available. The proposed algorithm is used for the activity detection of various sources, each with different characterization in terms of the movements in the video signal and in variations in the spectrum of the audio signal. Experimental results demonstrate the advantage and significance of including a video signal for the activity detection of sound sources.

ACKNOWLEDGMENT

The authors would like to thank the associate editor and the anonymous reviewers for their constructive comments and useful suggestions.

REFERENCES

[1] D. Dov, R. Talmon, and I. Cohen, "Kernel method for voice activity detection in the presence of transients," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 12, pp. 2313–2326, Dec. 2016.
[2] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, 2001.
[3] R. Talmon, I. Cohen, S. Gannot, and R. 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.
[4] P. Somervuo, A. Härmä, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2252–2263, Nov. 2006.
[5] M. Cristani, M. Bicego, and V. Murino, "Audio-visual event recognition in surveillance video sequences," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 257–267, Feb. 2007.
[6] S. Chu, S. Narayanan, and C. J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1142–1158, Aug. 2009.
[7] H. D. Tran and H. Li, "Sound event recognition with probabilistic distance SVMs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1556–1568, Aug. 2011.
[8] X. Valero and F. Alias, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1684–1689, Dec. 2012.
[9] J. Dennis, H. D. Tran, and E. S. Chng, "Image feature representation of the subband power distribution for robust sound event classification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 367–377, Feb. 2013.
[10] I-H. Jhuo *et al.*, "Discovering joint audio–visual codewords for video event detection," *Mach. Vis.Appl.*, vol. 25, no. 1, pp. 33–47, 2014.
[11] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2015, pp. 1–7.
[12] G. Lafay, M. Lagrange, M. Rossignol, E. Benetos, and A. Roebel, "A morphological model for simulating acoustic scenes and its application to sound event detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 10, pp. 1854–1864, Oct. 2016.

[13] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," Tech. Rep. DCASE2016 Challenge, Sep. 2016.

[14] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Appl. Sci.*, vol. 6, no. 6, 2016, Art. no. 162.

[15] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *Proc. 2016 IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2016, pp. 6440–6444.

[16] S. Adavanne, G. Parascandolo, P. Pertilä, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," Tech. Rep. DCASE2016 Challenge, Sep. 2016.

[17] S. Sigtia, A. M. Stark, S. Krstulović, and M. D. Plumbley, "Automatic environmental sound recognition: Performance versus computational cost," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2096–2107, Nov. 2016.

[18] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct. 2015.

[19] J. W. Fisher, III, T. Darrell, W. T. Freeman, and P. A. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 772–778.

[20] G. Iyengar, H. J. Nock, and C. Neti, "Audio-visual synchrony for detection of monologues in video archives," in *Proc. 2003 IEEE Int. Conf. Multimedia Expo.*, 2003, vol. 1, pp. I-329–332.

[21] E. Kidron, Y. Schechner, and M. Elad, "Cross-modal localization via sparsity," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1390–1404, Apr. 2007.

[22] A. L. Casanovas and P. Vandergheynst, "Audio-based nonlinear video diffusion," in *Proc. 2010 IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2010, pp. 2486–2489.

[23] N. Seichepine, S. Essid, C. Fvotte, and O. Capp, "Soft nonnegative matrix co-factorizationwith application to multimodal speaker diarization," in *Proc. 2013 IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 3537–3541.

[24] N. Seichepine, S. Essid, C. Févotte, and O. Cappé, "Soft nonnegative matrix co-factorization," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5940–5949, Nov. 2014.

[25] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[26] M. Balasubramanian, E. L. Schwartz, J. B. Tenenbaum, V. de Silva, and J. C. Langford, "The isomap algorithm and topological stability," *Science*, vol. 295, no. 5552, pp. 7–7, 2002.

[27] M.l Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.

[28] D. L. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," *Proc. Nat. Acad. Sci.*, vol. 100, no. 10, pp. 5591–5596, 2003.

[29] R. R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 21, no. 1, pp. 5–30, 2006.

[30] D. Zhou and C. J. C. Burges, "Spectral clustering and transductive learning with multiple views," in *Proc. ACM 24th Int. Conf. Mach. Learn.*, 2007, pp. 1159–1166.

[31] M. B. Blaschko and C. H. Lampert, "Correlational spectral clustering," in *Proc. IEEE Conf. Comput. Vis.Pattern Recognit.*, 2008, pp. 1–8.

[32] V. R. De Sa, P. W. Gallagher, J. M. Lewis, and V. L. Malave, "Multi-view kernel construction," *Mach. Learn.*, vol. 79, no. 1, pp. 47–71, 2010.

[33] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1413–1421.

[34] A. Kumar and H. Daumé, "A co-training approach for multi-view spectral clustering," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 393–400.

[35] Y. Y. Lin, T. L. Liu, and C. S. Fuh, "Multiple kernel learning for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1147–1160, Jun. 2011.

[36] B. Wang, J. Jiang, W. Wang, Z. H. Zhou, and Z. Tu, "Unsupervised metric fusion by cross diffusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2997–3004.

[37] H. C. Huang, Y. Y. Chuang, and C. S. Chen, "Affinity aggregation for spectral clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 773–780.

[38] B. Boots and G. Gordon, "Two-manifold problems with applications to nonlinear system identification," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 623–630.

[39] M. M. Bronstein, K. Glashoff, and T. A. Loring, "Making Laplacians commute," arXiv:1307.6549, 2013.

[40] O. Lindenbaum, A. Yeredor, M. Salhov, and A. Averbuch, "Multiview diffusion maps," arXiv:1508.05550, 2015.

[41] R. R. Lederman and R. Talmon, "Learning the geometry of common latent variables using alternating-diffusion," *Appl. Comput. Harmon. Anal.*, 2015.

[42] D. Dov, R. Talmon, and I. Cohen, "Kernel-based sensor fusion with application to audio-visual voice activity detection," *IEEE Trans. Signal Process.*, vol. 64, no. 24, pp. 6406–6416, Dec. 2016.

[43] R. Talmon and H. Wu, "Latent common manifold learning with alternating diffusion: Analysis and applications," arXiv:1602.00078, 2016.

[44] E. J. Ong and R. Bowden, "Robust lip-tracking using rigid flocks of selected linear predictors," in *Proc. 8th IEEE Int. Conf. Automat. Face Gesture Recognit.*, 2008.

[45] 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.

[46] Q. Liu, W. Wang, and P. Jackson, "A visual voice activity detection method with adaboosting," in *Proc. IET Sensor Signal Process. Defence*, 2011, pp. 1–5.

[47] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[48] T. Michaeli, W. Wang, and K. Livescu, "Nonparametric canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016.

[49] 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.

[50] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[51] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. 1st Int. Conf. Music Inf. Retrieval*, 2000.

[52] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ASR2000-Automat. Speech Recognit., Challenges New Millenium ISCA Tuts Res. Workshop*, 2000, pp. 18–20.

[53] [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[54] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[55] 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.

[56] 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.

[57] 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.

[58] D. Dov, R. Talmon, and I. Cohen, "Audio-visual voice activity detection using diffusion maps," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, pp. 732–745, Apr. 2015.

[59] [Online]. Available: http://www.freesound.org.

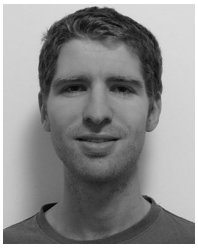
[60] D. Dov, R. Talmon, and I. Cohen, "Kernel method for speech source activity detection in multi-modal signals," in *Proc. IEEE Int. Conf. Sci. Elect. Eng.*, Nov. 2016, pp. 1–5.

[61] S. Tamura, M. Ishikawa, T. Hashiba, T. Shin'ichi, and S. Hayamizu, "A robust audio-visual speech recognition using audio-visual voice activity detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 2694–2697.

[62] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic-phonetic continous speech database," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Feb. 1993.

[63] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe, "Analyzing free-standing conversational groups: A multimodal approach," in *Proc. 23rd ACM Int. Conf. Multimedia*, New York, NY, USA, 2015, pp. 5–14.

[64] D. Kushnir, "Active-transductive learning with label-adapted kernels," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, 2014, pp. 462–471.



David Dov received the B.Sc. (*summa cum laude*) and M.Sc. (*cum laude*) degrees in electrical engineering in 2012 and 2014, respectively, from Technion—Israel Institute of Technology, Haifa, Israel, where he is currently working toward the Ph.D. degree in electrical engineering.

From 2010 to 2012, he was 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 geometric methods for data analysis, multisensors signal processing, speech processing, and multimedia.

He received the IBM Ph.D. Fellowship for 2016–2017, the Jacobs Fellowship for 2014, the Excellence in Teaching Award for outstanding teaching assistants in 2013, the Meyer Fellowship, the Cipers Award and the Finzi Award for 2012, the Wilk Award for excellent undergraduate project from the SIPL, Electrical Engineering Department, Technion for 2012, and Intel Award for excellent undergraduate students for 2009.

Ronen Talmon received the B.A. degree (*cum laude*) in mathematics and computer science from Open University in 2005, and the Ph.D. degree in electrical engineering from Technion—Israel Institute of Technology, Haifa, Israel, in 2011.

From 2000 to 2005, he was a Software Developer and a 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 2013, he was a Gibbs Assistant Professor in the Mathematics Department, Yale University, New Haven, CT, USA. In 2014, he joined the Department of Electrical Engineering of the Technion, where he is currently an Assistant Professor of electrical engineering.

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

He received 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) received the B.Sc. (*summa cum laude*), M.Sc., and Ph.D. degrees in electrical engineering from Technion—Israel Institute of Technology, Haifa, Israel, 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 Technion—Israel Institute of Technology, where he is currently a Professor of electrical engineering. 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. He was a 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 include 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 received 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. He was an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and the IEEE SIGNAL PROCESSING LETTERS, and a member of the IEEE Speech and Language Processing Technical Committee.