Sequential Audio-visual Correspondence With Alternating Diffusion Kernels

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

***Abstract*—A fundamental problem in multi-modal signal processing is to quantify relations between two different signals with respect to a certain phenomenon. In this paper, we address this problem from a kernel-based perspective and propose a measure that is based on affinity kernels constructed separately in each modality. This measure is motivated from both a kernel density estimation point of view of predicting the signal in one modality based on the other, as well as from a statistical model, which implies that high values of the proposed measure are expected when signals highly correspond to each other. Considering an online setting, we propose an efficient algorithm for the sequential update of the proposed measure, and demonstrate its application to eye-fixation prediction in audio-visual recordings. The goal is to predict locations within a video recording at which people gaze when watching the video. As studies in psychology imply, people tend to gaze at the location of the audio source, so that their prediction becomes equivalent to locating the audio source within the video. Therefore, we propose to predict eye-fixations as regions within the video with the highest correspondence to the audio signal, thereby demonstrating the improved performance of the proposed method.**

I. Introduction

Fusion of multi-modal signals, i.e., signals measured in multiple sensors of different types, has recently attracted a considerable attention in the signal processing and data analysis communities. In this paper, we consider a particular aspect of the fusion problem addressing the question: to what extent signals from different modalities correspond to each other. We regard to the *correspondence* as the level at which two signals measure the same source or phenomenon. A challenging example we consider is the correspondence between audio and video signals, which may be useful for the analysis of audio-visual sound scenes. For example, regions within the video having high levels of correspondence to the audio signal comprise the location of the audio source as we show in this paper.

We consider the correspondence between multi-modal signals from a kernel-based geometric perspective, also termed manifold learning. Such kernel-based approaches were originally designed for the analysis of single-modal signals [1]–[5]. They are usually based on the construction of an affinity kernel capturing similarities (relations) between samples of the signal, followed by an eigenvalue decomposition to obtain a low dimensional representation. In the past decade, several studies investigated the extension of these methods to the multi-modal case by exploiting different combinations of affinity kernels constructed separately for each modality [6]–[19]. These studies, however, focus on a different problem of how to obtain a unified representation of the multi-modal signals rather than the correspondence between them.

A fusion approach that is based on a product of affinities kernel was studied in [17]–[19]. Lederman and Talmon analyzed the kernel product in a continuous setting, showing that it recovers the common components from multi-modal observations. In [19], we have studied the kernel product from a graph theoretic point of view and proposed a method for the selection of the kernel bandwidth. In addition, Michaeli et al. showed in [18] the equivalence of this fusion approach to a non-parametric variant of kernel canonical correlation analysis (CCA).

Here we consider the correspondence between multi-modal signals, which was not previously addressed in [6]–[17], [19]. Furthermore, we address the problem of an online setting. By design, kernel methods are memory consuming since for a signal comprising N samples, they require a construction of an affinity kernel of size $N \times N$. In addition, the computational cost of the eigenvalue decomposition in these methods is very high. Accordingly, kernel methods are often constructed only from part of the available data [18], [20], and then extended to other samples using, e.g., Nyström method [21]. In this context, we mention the studies presented in [22]–[24], which examined adaptation of kernel methods over time in the single-modal case. However, these studies mainly focus on efficient computation of eigenvectors over time, which is not addressed here.

As an application of the correspondence between multi-modal signals, we consider the problem of eye-fixation prediction in audio-visual recordings. Eye-tracking experiments in psychology imply that people tend to gaze at the locations of sound sources within video recordings [25]–[32]. Accordingly, the localization of the audio source within the video is the main component in the prediction of eye-fixations as we show in this paper. Izadinia et al. [33] addressed this problem by exploiting canonical correlations between the audio signal and regions of the video, which were segmented in advance using a video-based approach. Min et al. [32] extended this framework by combining audio-visual correlations with cues, which are merely based on the video signal, for eye-fixation prediction. Zhang et al. [34], proposed to map audio-visual data into an embedded domain constructed using kernel CCA with multiple kernels. Then, they used the distance between audio-visual data in this domain as a measure of correspondence for the task of content retrieval.

The authors are with the Viterbi Faculty 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).

This research was supported by the Israel Science Foundation (grants no. 576/16 and 1490/16), and the ISF-NSFC joint research program (grant No. 2514/17).

The problem of audio-visual localization was also addressed in [35]–[38], typically formulated as an optimization problem for learning unified representations of the audio-visual signals. For example, Kidron et al. [36] extended the framework of CCA by introducing a regularization term based on the sparsity of events in which the audio and visual signals are correlated. Based on the solution of the associated optimization problems, the methods presented in these studies are computationally expensive, which restricts their applicability in an online setting.

We further note here the studies presented in [39], [40], which considered signals obtained in multi-channel microphone arrays, in addition to the video camera. In our study, however, we focus on measuring the correspondence between two modalities of signals obtained in a video camera and a *single* microphone. In addition, we note [41], in which the authors proposed to train a neural network for speaker detection, and more recent approaches for multimodal fusion via deep learning termed deep CCA [42]. Deep learning based methods such as [42], [43] are typically trained on large datasets. To the best of our knowledge, large datasets are not available for the task of eye-fixation prediction, and methods based on deep learning were not applied to this task.

A different variant of the problem of correspondence is further studied in the computer graphics community, where the goal is to match between pairs of points from two sets corresponding to two different shapes. Interestingly, the kernel product was recently used to address this variant of the correspondence problem in [44], [45]; Vestner et al. [45] formulated a linear assignment problem, in which finding the assignments of the pairs is equivalent to rearranging rows of the kernels of each set ("modality") prior to their product.

In this paper, we propose a measure of correspondence between multi-modal signals based on the trace of the kernel product. We show how variants of this measure naturally arise in the context of kernel density estimation, studied in [18] and [45]. In addition, we analyze this measure from a graph theoretic point of view using the statistical model we presented in [19] for describing the connectivity of graphs corresponding to the different modalities. We show that the higher the trace of the kernel product the higher is the correspondence between the multi-modal signals. Then, we show how to efficiently update this measure in an online setting for new incoming samples. Finally, we demonstrate the performance of the proposed measure for localization of audio sources in video and for prediction of eye fixations on a dataset recently presented in [32]. The proposed measure not only outperforms competing methods, but also allows to process the videos in a sequential manner. In addition, it allows to reduce the weight of other cues, based only on video, for the prediction of eye fixations implying the strong relation between the audio signal and eye fixations.

The remainder of the paper is organized a follows. In Section II, we review the construction of the kernel product and its use for sensor fusion. The analysis of the proposed measure for multi-modal correspondence from kernel density estimation perspective and from a graph point of view, and its online computation are present in Section III. In Section IV, we analyze the complexity of the proposed measure. Finally, in Section V, we demonstrate applications of audio-visual localization and eye-fixation prediction.

II. Review of The Kernel Product For Multi-modal Sensor Fusion

Let $\{(\mathbf{v}_n, \mathbf{w}_n)\}_{n=1}^{N}$ be a set of N pairs of data-points measured by two different sensors, where $\mathbf{v}_n \in \mathbb{R}^{L_v}$ and $\mathbf{w}_n \in \mathbb{R}^{L_w}$ are some feature representations of the nth time frame of the first and the second modalities, respectively. In the context of eye-fixation prediction, these are the audio and the video features representing the nth video frame, where we assume that the audio signal is processed in frames, which are aligned to the video signal. The fusion process between the two modalities is based on the construction of affinity kernels, $\mathbf{K}_v \in \mathbb{R}^{N\times N}$ and $\mathbf{K}_w \in \mathbb{R}^{N\times N}$, one for each modality. The (n,m)th entry of $\mathbf{K}_v \in \mathbb{R}^{N\times N}$, denoted by $K_v(n,m)$, is given by:

$$K_v(n,m) = \exp\left(-\frac{\|\mathbf{v}_n - \mathbf{v}_m\|^2}{\epsilon_v}\right), \tag{1}$$

where $\|\cdot\|$ is the L_2 norm, ϵ_v is a scaling parameter, and $\mathbf{K}_w \in \mathbb{R}^{N\times N}$ is defined similarly[1]. We denote by $\mathbf{M}_v \in \mathbb{R}^{N\times N}$ a normalized version of $\mathbf{K}_v$, given by:

$$\mathbf{M}_v = \mathbf{D}_v^{-1}\mathbf{K}_v, \tag{2}$$

where $\mathbf{D}_v \in \mathbb{R}^{N\times N}$ is a diagonal matrix whose (n,n)th element is the sum of the nth row of $\mathbf{K}_v$. The two modalities are fused via the product between the normalized kernels, $\mathbf{M}_v\mathbf{M}_w$, which is referred to as the unified kernel.

Due to the normalization, $\mathbf{M}_v$ and $\mathbf{M}_w$ are both row stochastic matrices, and so is the unified kernel. The continuous counterparts of these three kernels have an interpretation involving diffusion processes. Specifically, the unified diffusion process is applied to the two modalities in an alternating manner, so that it is referred to as "alternating diffusion maps" [17], [46]. When applied to a certain modality, the unified diffusion process attenuates factors specific to other modalities, which are often considered interferences, justifying its use for the representation of multi-modal signals.

III. Kernel-based Measure For Multi-modal Correspondence

We propose to use the trace of the kernel product as a measure of correspondence between multi-modal signals in an online setting. By revisiting [18] and [45], we discuss in Subsection III-A variants of the proposed measure in the context of kernel density estimation. In Subsection III-B, we present a new interpretation of this measure using a statistical model arising from a graph interpretation of the kernels. Finally, we present an efficient algorithm for the online calculation of the proposed measure in Subsection III-C.

[1]All entities related to the first and the second modalities are denoted in the paper by the subscripts or superscripts v and w, respectively. If not explicitly stated, the entities of the second modality are defined throughout the paper similarly to the first modality.

A. From the Perspective of Kernel Density Estimation

The study presented in [45] addressed the problem of matching between pairs in the sets $\{\mathbf{v}_n\}_{n=1}^N$ and $\{\mathbf{w}_n\}_{n=1}^N$, assuming that the true match between a subset of $\tilde{N}$ pairs is available in advance and that the other points are given in a random order. This problem arises in computer graphics applications, where one is interested in matching between two shapes, each discretized by N points, such that the shapes correspond to the sets $\{\mathbf{v}_n\}_{n=1}^N$ and $\{\mathbf{w}_n\}_{n=1}^N$. The authors proposed to match between the pair (v, w) in the continuous setting by finding a mapping $w = g(v)$ such that $g(v)$ is estimated by:

$$\hat{g}(v) = \arg\max_w f(v, w),$$

where f is the joint density of the pair. Namely, the mapping is obtained by the MAP estimator of one view by the other. The joint density is estimated via the kernel density estimation framework:

$$f(v, w) \propto \sum_{n=1}^{\tilde{N}} \exp\left(-\frac{\|v - v_n\|^2}{\epsilon_v}\right) \exp\left(-\frac{\|w - w_n\|^2}{\epsilon_w}\right).$$

The authors considered a discretization leading to the following optimization problem:

$$\arg\max_{\mathbf{P}} \mathrm{Tr}\left\{\mathbf{K}_v \mathbf{K}_w^T \mathbf{P}\right\}, \tag{3}$$

where $\mathbf{K}_v, \mathbf{K}_w \in \mathbb{R}^{N \times \tilde{N}}$ are defined similarity to (1) and $\mathbf{P} \in \mathbb{R}^{N \times N}$ is an assignment matrix, whose (n, m)th entry equals one if points $\mathbf{v}_n$ and $\mathbf{w}_m$ match and zero otherwise.

In our case, the two sets are aligned, i.e., $\mathbf{v}_n$ and $\mathbf{w}_n$ match to each other since they are samples taken at the same time n. We hence expect the optimal solution $\mathbf{P}$ be the identity matrix and the highest correspondence value is the trace of the kernel product. Namely, the kernel product calculated over the aligned set yields the highest correspondence value compared to a kernel product constructed based on any other permutation between the data-points.

Michaeli et al. studied in [18] the kernel density estimation of $f(v, w) / (f(v) f(w))$, where $f(v)$ and $f(w)$ are the densities of the data in the two modalities. They interpreted this density as the MMSE estimator of the data in one modality based on the other. They showed that the corresponding discretized operator is the kernel product:

$$\mathbf{M} = \mathbf{M}_v \mathbf{M}_w^T, \tag{4}$$

so that it can replace the kernel $\mathbf{K}_v \mathbf{K}_w^T$ in (3) for the assignment problem. In addition, they showed that the singular value decomposition of $\mathbf{M}$ maximizes the linear correlation between the views in a specifically designed kernel space such that the method may be considered as a variant of kernel CCA. Let $\sigma_1, \sigma_2, ..., \sigma_N$ be the singular values of $\mathbf{M}$, and let $\boldsymbol{\sigma} \in \mathbb{R}^N$ be a vector, whose ith element is σ_i. According to [18], the correlation is given by the sum of the singular values, which is the l_1 norm of $\boldsymbol{\sigma}$, namely $\|\boldsymbol{\sigma}\|_1 = \sum_{n=1}^N |\sigma_n|$. Note that the eigenvalues of $\mathbf{M}\mathbf{M}^T$ are the squares of the singular values of $\mathbf{M}$, i.e., σ_i^2. Conceivably, using [18] but with the different l_2 norm results in $\|\boldsymbol{\sigma}\|_2^2 = \sum_{n=1}^N |\sigma_n|^2$, which is nothing but the trace of $\mathbf{M}\mathbf{M}^T$.

We note that we found in our experiments that the different variants of the measure of correspondence perform similarly. Here, we propose to use the trace of the kernel product $\mathbf{M}$ as a measure of correspondence between multi-modal signals, since it allows us to design an efficient algorithm for an online update of its trace.

We further note in this context the Hilbert-Schmidt independence criterion (HSIC) as a related measure of correspondence. The HSIC is a statistical criterion which measures independence between the modalities based on the Hilbert-Schmidt norm [47]. Similarly to the proposed measure and assuming that the data is centered, the HSIC is estimated by the trace of the product $\mathbf{K}_v \mathbf{K}_w$. This measure, however, does not have the interpretation of a diffusion process and has inferior performance as we show in Section V.

B. Statistical Interpretation

In this subsection, a statistical interpretation of the measure $\mathrm{Tr}\{\mathbf{M}\}$ from a graph theory point of view is presented. The affinity kernel $\mathbf{K}_v$ in (1) defines a graph, whose vertices correspond to the N data-points, and the weights of the edges are given by $K_v(n, m) = \exp\left(-\frac{\|\mathbf{v}_n - \mathbf{v}_m\|^2}{\epsilon_v}\right)$ (1). Points n and m are considered connected if $\|\mathbf{v}_n - \mathbf{v}_m\|^2 < \epsilon_v$ such that high affinities are obtained between them, and they are disconnected when $\|\mathbf{v}_n - \mathbf{v}_m\|^2 > \epsilon_v$, so that the affinity between them is negligible. While these considerations were used in [19], [48] for the selection of the kernel bandwidth ϵ_v, we utilize them for the analysis of the proposed measure.

We encode the connectivity between points n and m using a simplified statistical model, which we presented in [19]. Let $\mathbb{I}_{n,m}^v$ denote an indicator which equals one if the pair of points (n, m) is connected and zero otherwise. Assuming that each pair is connected with probability p_v independently from all other pairs, we have that:

$$\mathbb{I}_{n,m}^v = \left\{ \begin{array}{ll} 1, & \text{w.p. } p_v \\ 0, & \text{otherwise} \end{array} \right\}, \tag{5}$$

so that the indicators $\left\{\mathbb{I}_{n,m}^v\right\}$ are independent and identically distributed.

We proceed to the normalized version of the kernel recalling that its (n, m)th entry is given by $M(n, m) = K(n, m) / D(n, n)$, where $D(n, n)$ is the sum of the nth row. According to the statistical model, each point is connected on average to $1 + p_v(N - 1)$ points for large values of N, where we assume that each point is connected to itself. Accordingly, we define a measure for the connectivity of the normalized kernel, $\left\{\mathbb{J}_{n,m}^v\right\}$, similarly to (5):

$$\mathbb{J}_{n,m}^v = \frac{1}{1 + p_v(N - 1)} \mathbb{I}_{n,m}.$$

We assume that the correspondence is related to the correlation between the indicators in the two modalities. The higher the correspondence between points n and m, the higher is the correlation between their measures $\mathbb{J}_{n,m}^v$ and $\mathbb{J}_{n,m}^w$. We

consider two extreme cases, in which the two modalities are uncorrelated or fully correlated, and calculate the expected value of the trace of the kernel product in these cases:

$$\begin{aligned}\mathbb{E}\left(\text{Tr}\{\mathbf{M}\}\right) &= \mathbb{E}\left(\text{Tr}\left\{M_v M_w^T\right\}\right) \\ &= \mathbb{E}\left(\sum_{n=1}^{N}\sum_{m=1}^{N} M_v\left(n,m\right) M_w\left(n,m\right)\right).\end{aligned}$$

When the two modalities are uncorrelated, we have:

$$\begin{aligned}\mathbb{E}\left(\text{Tr}\{\mathbf{M}\}\right) &= N^2\mathbb{E}\left(M_v\left(n,m\right)\right)\mathbb{E}\left(M_w\left(n,m\right)\right) \\ &= N^2\mathbb{E}\left(\mathbb{J}_{n,m}^{v}\right)\mathbb{E}\left(\mathbb{J}_{n,m}^{w}\right) = N^2\frac{p_v}{1+p_v\left(N-1\right)}\frac{p_w}{1+p_w\left(N-1\right)}.\end{aligned}$$

On the other hand, when the correlation between the views is maximal, we have:

$$\begin{aligned}\mathbb{E}\left(\text{Tr}\{\mathbf{M}\}\right) &= N^2\mathbb{E}\left(M_v^2\left(n,m\right)\right) \\ &= N^2\mathbb{E}\left(\mathbb{J}_{n,m}^{v}\right)^2 = N^2\frac{p_v}{\left(1+p_v\left(N-1\right)\right)^2},\end{aligned}$$

where we assumed that $p_v = p_w$. As a result, there is a factor of $p_v \in (0,1)$ between the two extremes implying that the trace of the kernel product is expected to receive higher values when the data in the two views correspond to each other.

C. Online Computation of the Multi-modal Measure of Correspondence

We propose an algorithm for an efficient update of the trace of the kernel product, $\text{Tr}\{\mathbf{M}\}$, in a frame by frame manner. Given a new incoming frame, whose time index is denoted by $N+1$, our goal is to efficiently calculate the trace of the kernel product corresponding to frames $2,3,...,N+1$ without recalculating the kernels of each modality and the product between them. Based on properties of the trace and the symmetry of the kernels $\mathbf{K}_v$ and $\mathbf{K}_w$, the following derivation shows that only the affinities between the new incoming frame and the other $N-1$ points are required to compute the trace.

Let $\mathbf{D} \in \mathbb{R}^{N\times N}$ and $\mathbf{K} \in \mathbb{R}^{N\times N}$ denote the products $\mathbf{D}_v^{-1}\mathbf{D}_w^{-1}$ and $\mathbf{K}_v\mathbf{K}_w$, respectively. Our main observation, presented in Proposition 1, implies that the trace of the kernel product can be expressed by the the diagonal elements of these two matrices, which in turn may be sequentially updated.

Proposition 1. *The trace of the kernel product is given by:*

$$Tr\left\{\mathbf{M}\right\} = \sum_{n=1}^{N} D(n,n)K(n,n) \tag{6}$$

Proof: We recall that the trace of the kernel product $\mathbf{M}$ is given by:

$$\begin{aligned}\text{Tr}\left\{\mathbf{M}\right\} = \text{Tr}\left\{\mathbf{M}_v\mathbf{M}_w^T\right\} &= \text{Tr}\left\{\mathbf{D}_v^{-1}\mathbf{K}_v\left(\mathbf{D}_w^{-1}\mathbf{K}_w\right)^T\right\} \\ &= \text{Tr}\left\{\mathbf{D}_v^{-1}\mathbf{K}_v\mathbf{K}_w^T\mathbf{D}_w^{-T}\right\}.\end{aligned}$$

Since both $\mathbf{K}_w$ and $\mathbf{D}_w$ are symmetric, we have:

$$\text{Tr}\left\{\mathbf{M}\right\} = \text{Tr}\left\{\mathbf{D}_v^{-1}\mathbf{K}_v\mathbf{K}_w\mathbf{D}_w^{-1}\right\}.$$

In addition, the trace is invariant to cyclic shift and $\mathbf{D}_v$, $\mathbf{D}_w$ are diagonal, so that we have:

$$\text{Tr}\left\{\mathbf{M}\right\} = \text{Tr}\left\{\mathbf{D}_v^{-1}\mathbf{D}_w^{-1}\mathbf{K}_v\mathbf{K}_w\right\}.$$

By substituting $\mathbf{D}$ and $\mathbf{K}$, we rewrite the last expression using the Hadamard (point-wise) product:

$$\text{Tr}\left\{\mathbf{M}\right\} = \sum_{n=1}^{N}\sum_{m=1}^{N} D(n,m)K(n,m).$$

Finally, since $\mathbf{D}$ is diagonal, we obtain (6). ∎

Next we show how to sequentially update (6) using merely the affinities to the new frame $N+1$, $K_v(n,N+1)$ and $K_w(n,N+1)$ for all $n \in \{2,3,...,N\}$. Let $\tilde{\mathbf{M}}$ be the kernel product calculated from frames $2,3,...,N+1$, whose trace is given by:

$$\text{Tr}\left\{\tilde{\mathbf{M}}\right\} = \sum_{n=1}^{N}\tilde{D}(n,n)\tilde{K}(n,n), \tag{7}$$

where $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{K}}$ are the updated versions of $\mathbf{D}$ and $\mathbf{K}$, respectively. By the law of matrix product, the term $\tilde{K}\left(n,n\right)$ is given by:

$$\tilde{K}(n,n) = \sum_{m=2}^{N+1} K_v(n,m)K_w(n,m), \tag{8}$$

so that it is sequentially updated by:

$$\begin{aligned}\tilde{K}(n,n) = K(n,n) - K_v(n,1)K_w(n,1) \\ +K_v(n,N+1)K_w(n,N+1).\end{aligned} \tag{9}$$

The term $\tilde{D}\left(n,n\right)$ is given by:

$$\tilde{D}\left(n,n\right) = \frac{1}{\tilde{D}_v\left(n,n\right)\tilde{D}_w\left(n,n\right)}, \tag{10}$$

where:

$$\tilde{D}_v\left(n,n\right) \triangleq \sum_{m=2}^{N+1} K_v\left(n,m\right). \tag{11}$$

Accordingly, $\tilde{D}\left(n,n\right)$ is calculated via a sequential update of $\tilde{D}_v\left(n,n\right)$ and $\tilde{D}_w\left(n,n\right)$:

$$\tilde{D}_v(n,n) = D_v(n,n) - K_v(n,1) + K_v(n,N+1). \tag{12}$$

We summarize the proposed algorithm for the efficient update of the kernel product in Algorithm 1.

IV. Complexity analysis and run-time simulation

We analyze the computational complexity of updating the trace of the kernel product according to Algorithm 1. Equation (7) requires N (scalar) multiplications, i.e., one multiplication $\tilde{D}(n,n)\tilde{K}(n,n)$ for each n, which are then followed by the sum over N, i.e., N scalar summations. The calculation of $\tilde{D}(n,n)$ for $n = 1,2,...,N$ requires according to (10) $2N$ operations, or more specifically, N multiplications $\tilde{D}_v(n,n)\tilde{D}_w(n,n)$ and N divisions. In turn, $\tilde{D}_v(n,n)$ and $\tilde{D}_w(n,n)$ are given according to (12) by three summations each, which gives a total of $6N$ summations. Finally, computing $\tilde{K}(n,n)$ in (9) for $n = 1,2,...,N$ requires $2N$ scalar multiplication and $3N$ summations. In summary, the update of the trace of the kernel product has the complexity of $O\left(N\right)$, and specifically, it requires $10N$ summations and $5N$ multiplications. We further note that in practice, we calculate (7), (9), (10) and (12) simultaneously for $n \in (1,2,...,N)$ by writing

Algorithm 1 Sequential update of the proposed measure for multi-modal correspondence

Initialization:

Input: a set of N pairs of data-points $\{(\mathbf{v}_n,\, \mathbf{w}_n)\}_{n=1}^{N}$
Output: $\mathbf{K}$ and $\mathbf{D}$

1: Calculate the affinity kernels $\mathbf{K}_v$ and $\mathbf{K}_w$ according to (1)
2: Calculate the normalization matrices $\mathbf{D}_v$, $\mathbf{D}_w$
3: Calculate $\mathbf{K} = \mathbf{K}_v\mathbf{K}_w$, $\mathbf{D} = \mathbf{D}_v^{-1}\mathbf{D}_w^{-1}$

Update:

Input: a new incoming pair $(\mathbf{v}_{N+1},\, \mathbf{w}_{N+1})$, $\mathbf{K}$ and $\mathbf{D}$
Output: $\mathrm{Tr}\left\{\tilde{\mathbf{M}}\right\}$
4: Calculate the affinities to the new pair according to (1): $K_v(n, N+1)$ and $K_w(n, N+1)$, $n \in (2, 3, ..., N)$
5: Update $\tilde{K}(n, n)$ according to (9)
6: Update $\tilde{D}_v(n, n)$, $\tilde{D}_w(n, n)$ according to (11)
7: Update $\tilde{D}(n, n)$ according to (12)
8: Update $\mathrm{Tr}\left\{\tilde{\mathbf{M}}\right\}$ according to (6)

Note: Steps 4-7 may be vectorized for simultaneous calculations of $n \in (1, 2, ..., N)$

them in a vectorized form, such that the only dependence on N is the update of the affinity kernels $K_v(n, N+1)$ and $K_w(n, N+1)$.

As a comparison, we consider the complexity of the calculation of the trace of $\mathbf{M}$ assuming that the matrices $\tilde{\mathbf{D}}_v, \tilde{\mathbf{D}}_w, \tilde{\mathbf{K}}_v, \tilde{\mathbf{K}}_w$ are efficiently updated. These matrices may be updated by removing their first row and columns and adding the new row and column, corresponding to the incoming frame. In this case, the updated kernel $\tilde{\mathbf{M}}$ is given by the multiplication between these four matrices, the complexity of which is $O\left(N^3\right)$ using naive matrix multiplication methods, and even when efficient algorithms are used, the complexity remains above $O\left(N^2\right)$. We relate to this alternative approach as "single modal update" since it was studied in [23] in the single-modal setting.

To demonstrate the run-time efficiency of Algorithm 1, we compare it to the alternative approach for the calculation of the proposed measure for multi-modal correspondence using synthetic data. In addition, we compare the proposed algorithm to a naive algorithm, in which, given a new incoming frame, the trace is computed from scratch. In the first experiment, we study the effect of the number of features in the dataset $\{(\mathbf{v}_n,\, \mathbf{w}_n)\}_{n=1}^{N}$, i.e., L_v, L_w, on the run-time. We run 100 simulations, sweeping in each simulation the number of features from 10 to 300. For all simulations, we set $N = 100$ and consider the update of the trace of the kernel product for 1000 new incoming frames.

The average run-time for the different number of features is presented in Fig. 1 (top). It can be seen that the run-time of the naive algorithm linearly increases with the number of features making it not practical for online applications. The bottleneck of the naive algorithm lies in the calculation of the affinity kernel $\mathbf{K}_v$ and $\mathbf{K}_w$, which are recomputed for each new incoming frame. In contrast, the proposed algorithm and the "single modal update" approach are barely affected by the increase in the number of features. This is because in these methods, only the affinities $\mathbf{K}_v(n, N+1)$, $\mathbf{K}_w(n, N+1)$, $n \in (2, 3, ..., N)$ are calculated for the incoming frame $N+1$ instead of the whole affinity matrices.

In the second experiment, we further explore the difference between the proposed algorithm and the "single modal update" approach by comparing their run-time versus the number of pairs, N, in the dataset. In addition, we compare the proposed approach to Singular Value Decomposition (SVD) to demonstrate the run-time improvement obtained by avoiding from singular/eigen-decomposition, which is a common step in the construction of kernel-based methods. We use a truncated (fast) version of SVD taken from "Scikit-learn", a python package for machine learning [49]. In addition, we compare the run-time of the proposed approach to an implementation of kernel CCA taken from [50].

We set the number of features in this experiment to $L_v = L_w = 200$ and present the results in Fig. 1 (bottom). Although the "single modal update" method outperforms both SVD and KCCA, its run-time increases with N since it is based on the multiplication of the matrices $\tilde{\mathbf{D}}_v, \tilde{\mathbf{D}}_w, \tilde{\mathbf{K}}_v, \tilde{\mathbf{K}}_w$, whose sizes are $N \times N$. In contrast, the number of pairs has almost no effect on the proposed algorithm and it performs significantly faster. We note in this context that there exist efficient versions of kernel CCA such as the one presented in [51]. These methods, however, focus on reducing the memory consumption during the processing of large datasets rather than on online processing as in this paper, and they merely approximate kernel CCA using pre-trained models.

V. Experimental Results

A. Audio Localization in Video

We demonstrate the proposed measure of multi-modal correspondence for the problem of audio localization in videos. In the first experiment, we use four video recordings of U.S. presidential debates, taken from YouTube[2]. In each recording, only one of the speakers is active and the goal is to localize the speaker (the sound source).

Each video recording has the length of 90 sec and the resolution of 720×1280 pixels, and it is processed in 29 fps. We divide the video into a grid of rectangular cells each of 40×40 pixels and consider each cell as a separate video stream. Accordingly, the problem of localization is transformed to finding streams with high levels of correspondence to the audio signal. Each video cell is represented by motion vectors, which are widely used for the representation of visual speech signals [52]–[54]. We use a block size of 10×10 pixels, and form a feature vector of size $L_w = 32$ by concatenating the horizontal and the vertical velocities of each block.

[2]link to the videos presented in Figs. 2 (a) and (b): https://www.youtube.com/watch?v=d4Tinv8DMBM, time intervals: $6^{'}:30^{''}-8^{'}:00^{''}$ and $8^{'}:30^{''}-10^{'}:00^{''}$
link to the videos presented in Figs. 2 (c) and (d): https://www.youtube.com/watch?v=hx1mjT73xYE, time intervals: $3^{'}:15-4^{'}:45^{''}$ and $4^{'}:55^{''}-6^{'}:25^{''}$

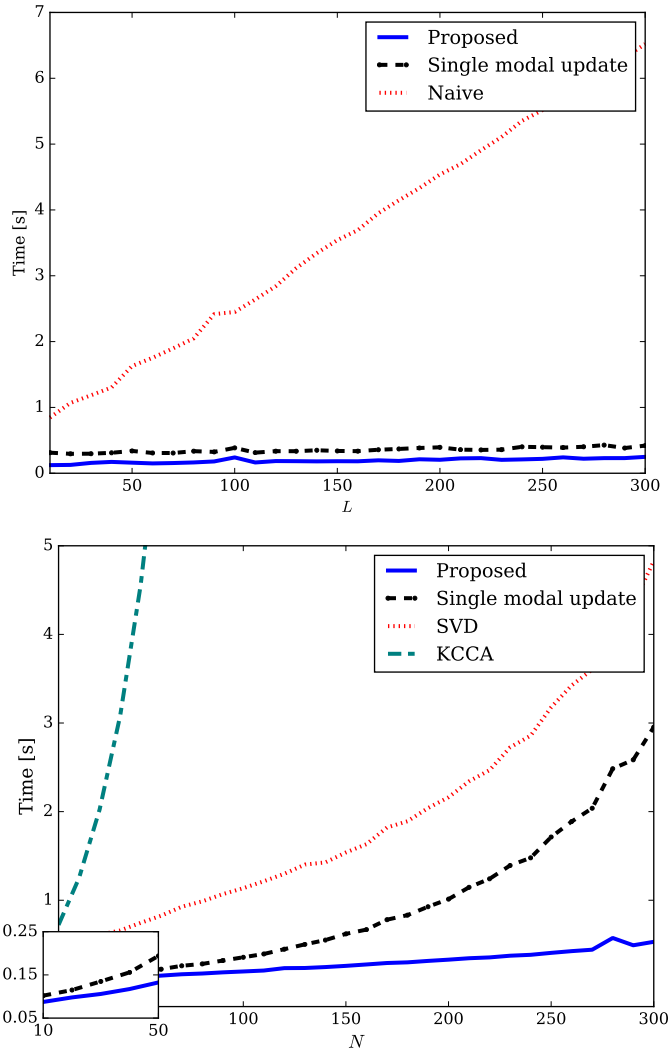


Fig. 1: Run-time of the algorithms for 1000 new incoming frames averaged over 100 simulations. (top) run-time versus the number of features $L = L_v = L_w$, the batch size is set to $N = 100$. (bottom) run-time versus batch size in frames N, the number of features is set to $L_v = L_w = 200$.

The audio signal is sampled at 44100 kHz and is processed with time frames of $\sim$ 66 ms with 50% overlap such that the audio and the video signals are aligned. We use 13 Mel-frequency cepstral coefficients (MFCCs) for the representation of each audio frame, i.e., the dimension of the audio signal is $L_v = 13$. The MFCCs are widely used for the representation of audio signals since they represent the spectrum of the signal in a compact form [55], and we have previously exploited them in [19].

We measure the correspondence between the audio signal and each one of the video streams using the proposed measure in (6) based on the product of kernels. We set the kernel bandwidths ϵ_v and ϵ_w according to [48] such that ϵ_v is given by:

$$\epsilon_v = C \max_m \left[\min_n \left(\|\mathbf{v}_n - \mathbf{v}_m\|^2 \right) \right].$$

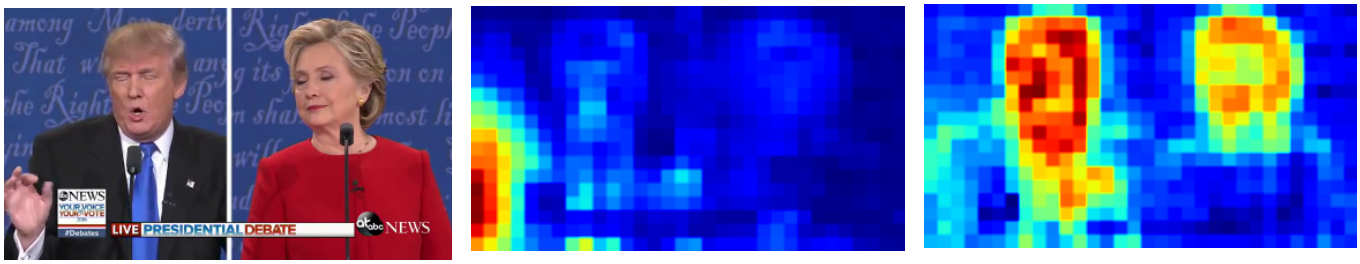
From a graph point of view, each point in the graph is connected when $C = 1$, and C is empirically set to the range of $2 - 3$ to guarantee connectivity of the graph. In [19], we analyzed the selection of the kernel bandwidth in the multi-modal case and showed that the graph which corresponds to the product of kernels remains connected even if C is chosen significantly smaller. Here for simplicity we set $C = 2$ and note that we found in our experiments that this value can be decreased without degrading the results.

We present in Fig. 2 (right column) the average levels of correspondence in the form of a heat map. It can be seen in the figures that streams located in the face region of the active speaker have high temperatures implying on large correspondence values to the audio signal. These findings coincide with previous studies linking between speech production and facial behavior [56]–[58]. Interestingly, the heat maps of Hillary Clinton and Donald Trump in Figs. 2 (a) and (b) indicate certain correspondence levels between the audio signal and the inactive speaker. However, this correspondence is significantly lower than the correspondence to the active speaker and it may be attributed to slight head movements, e.g., nodding with the head when listening to the active speaker. We also present in Fig. 2 (center column) heat maps obtained by averaging over the motion vectors in the videos over time. It can be seen that the level of movement obtained in the bottom left corners of Figs. 2 (a) and (b) is significantly higher compared to the face region of the active speaker. The movements of the hands are not related to the audio signal and they are considered strong interferences. From the perspective of alternating diffusion [17], [46], the proposed measure attenuates sensor specific factors, i.e., the movement of hands, allowing to successfully measure correspondence between the modalities.

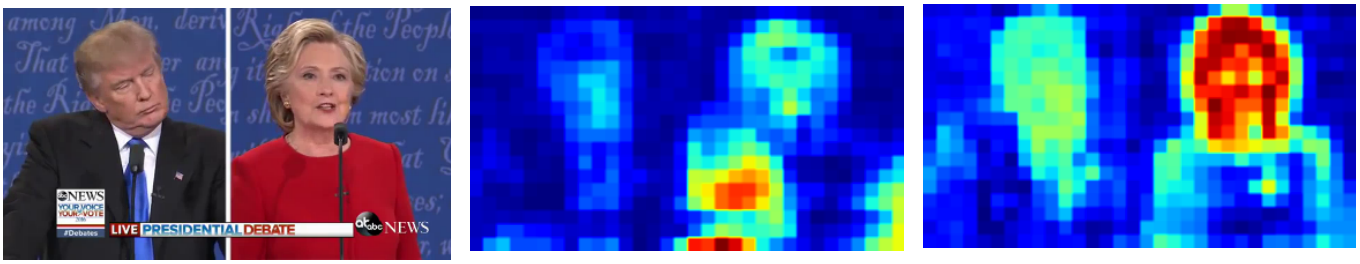
B. Eye-fixation Prediction

We proceed to the second experiment, where we apply the proposed measure for multi-modal correspondence to the problem of eye-fixation prediction. We use a dataset of 45 videos of lengths $5 - 10$ s, recently presented in [32]. The videos consist of different natural scenes such as people speaking or playing different types of musical instruments. The true eye fixations are collected using Tobii T120 Eye Tracker, which has a $17-$inch screen with the resolution of 1280×1024 pixels. The apparatus collects eye-fixations of 16 subjects watching each one of the videos. Accordingly, the eye-fixation data comprises binary images corresponding to the video frames such that a pixel in the image has the value of 1 if one of the subjects gazed at its location in the corresponding video frame and zero otherwise. The goal of the experiment is to predict the locations of the eye-fixations based solely on the audio and the video recordings. For more details regarding the dataset, we refer the reader to [32].

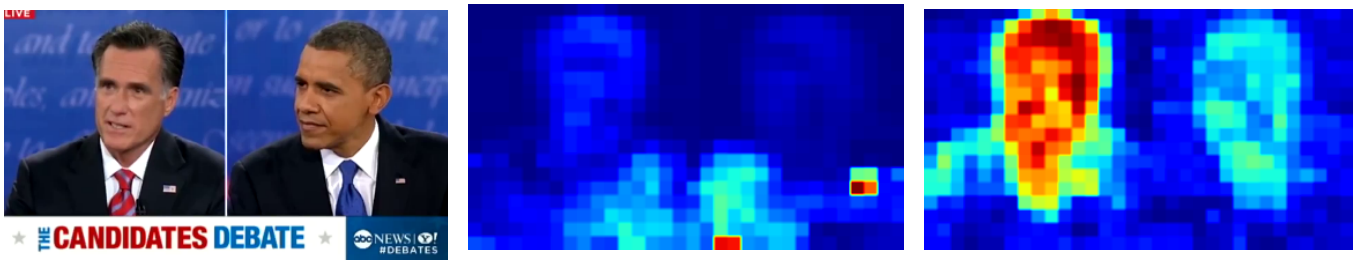
We compare the performance of the proposed measure of multi-modal correspondence to the method presented in [32]. The method is based on the representation of the audio and the video signals using MFCCs and motion vectors, respectively, similarly to the first experiment. Specifically, 10 MFCCs and 10 delta-MFCCs are used for the representation of the audio



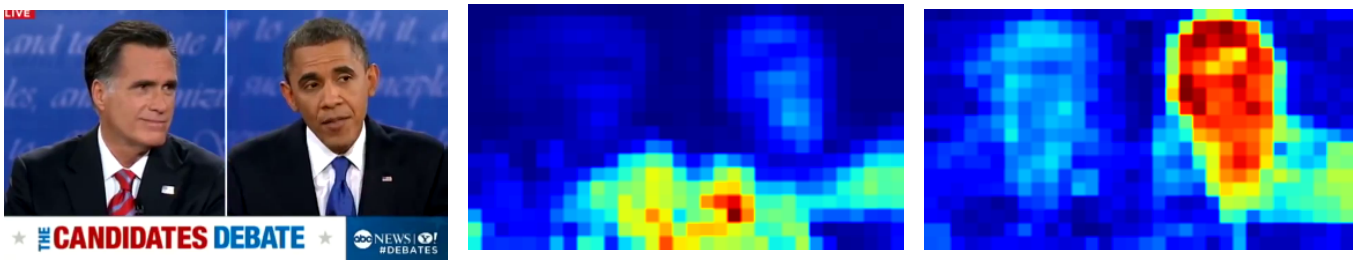
(a)



(b)



(c)



(d)

Fig. 2: Audio localization in video. Each video has the length of 90 sec, resolution of 720×1280 and frame rate of 29 fps. Left column:original image; center column: a heat map obtained by averaging on the motion vectors; right column: a heat map obtained by the proposed measure of correspondence. (a,c) Left speaker is active. (b,d) Right speaker is active.

signal. The video signal is first divided into super-voxels using a graph-based streaming segmentation method [59]. Then, each super-voxel is represented by the variance of its motion and acceleration, where the latter is the difference between the motion of the current frame with respect to the next frame and the motion between the current and the previous frames. The audio-visual correspondence is finally obtained by applying CCA such that the predicted regions are those related to the super-voxels with the maximal correlation to audio. Since in addition to the audio source, eye-fixations are also related to salient spatial and temporal events, the authors incorporate also cues which are based merely on the video signals. They generate, for each frame, a prediction map, based on the magnitude of the motion vectors. In addition, they compare between different state-of-the-art spatial saliency maps computed separately for each frame. Here, we choose the method presented in [60], which is based on computing a spectral residual of an image, since it was found to perform well in [32]. Finally, the three maps, related to the audio-visual correspondence and the spatial and the temporal cues are fused with equal weights using a simple sum. For more implementation details we refer the reader to [32].

Similarly to [32], we use three common measures to evaluate the prediction of eye fixations. First is the shuffled area under the curve (sAUC), in which receiver operating characteristic (ROC) curves are generated by sweeping a threshold between the minimal and maximal values of the saliency map. Since there are only a small number of true eye fixations in each frame, false locations are randomly shuffled from the (true) fixations in all other frames, such that there is an equal number of true and false pixels. Second is (linear) correlation coefficient (CC) obtained by calculating a two dimensional correlation between the estimated and the true fixation maps, where the latter is convolved with a Gaussian kernel. Last is the normalized scanpath saliency (NSS) score presented in [61]. NSS is the mean value of the predicted map at the true fixations, for which the predicted map is normalized to have a zero mean and a unit variance.

To apply the proposed measure for multi-modal correspondence for eye-fixation prediction, we use the same visual features as in [32]. We create an audio-visual correspondence map by calculating the correspondence between the audio features and the features of each one of the super-voxels, assigning the correspondence values to their corresponding pixels. Similarly to [32], we apply spatio-temporal smoothing to the audio-visual correspondence map and incorporate the other two maps, based on spatial and the temporal cues, respectively.

In Fig. 3, we present the performance of the proposed measure of correspondence for different values of the batch size N. The proposed measure is based on relations between geometric structures of the two modalities as they are encoded by the affinity kernels so that the number of frames has to be large enough to capture these structures. Conversely, the use of a too large number of frames may blur the estimate of eye-fixations since they change over time. The silver-lining of the trade-off is obtained in Fig. 3 for $N = 25$, a value which we use for comparison to the other methods. Figure 3 further implies that N has a small effect on the performance of the proposed measure, which, for a wide value range of N, outperforms the competing methods, whose performances are reported in Table I.

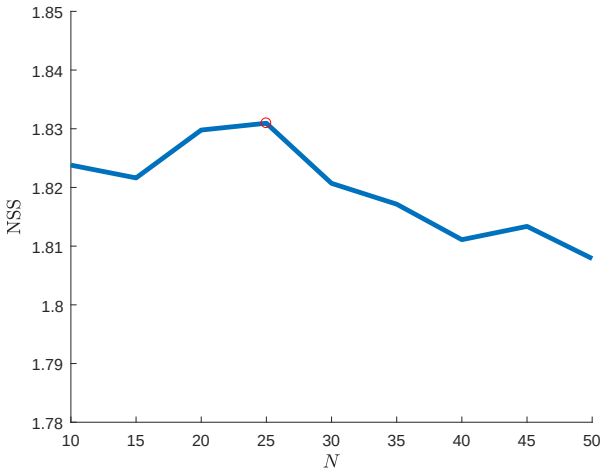


Fig. 3: The performance of the proposed measure of correspondence for eye-fixations prediction in terms of NSS versus N, the batch size in frames (blue solid line). Best performance obtained for $N = 25$ (red circle).

In Fig. 4, we present eye-fixation predictions obtained by the proposed algorithm in the form of heat maps. In addition to [32], we also compare the proposed method to kernel CCA. Figure 4 (a) comprises an example of a video frame of a person playing a piano such that the true eye fixations are centered at the region of the hands and the center of the body of the pianist. The maps, predicted by the proposed measure of multi-modal correspondence, as well as by kernel CCA, successfully indicate a high level of correspondence between the movement of the hands and the music. In contrast, the map, predicted by the method in [32], wrongly predicts the walking women in the background. Similarly, both [32] and kernel CCA wrongly predict as salient the movements in the background in Figs. 4 (b-d) and the arms movements of the player in Fig. 4 (e), which are not directly related to the production of sound. These movements may be considered as interferences, and they are properly attenuated by the proposed measure.

We further compare the proposed approach to other competing methods and present the results in Table I. We consider kernel CCA, the empirical HSIC and the method presented in [34] as alternative approaches for measuring correspondence between the audio and video modalities. The method in [34] is based on the use of kernel CCA with multiple kernels and suggests to measure correspondence according to the average distance between audio and video in the space of the kernels. We also compare the proposed method to [33], which is similar to [32] but only employs the audio-visual correspondence for eye-fixation prediction and does not use video-only based ques. Finally, we consider a variant of the method in [32] based only on the video signal such that only the spatial and temporal cues are used for the prediction of eye-fixation.

TABLE I: Comparison of the eye-fixation prediction scores.

Algorithm	sAUC	CC	NSS
Video only	0.7292	0.3612	1.4295
KCCA	0.7628	0.4362	1.7904
Empirical HSIC	0.7530	0.4197	1.7229
Zhang et al. 2016	0.7235	0.3725	1.4667
Izadinia et al. 2013	0.6915	0.3519	1.5165
Min et al. 2016	0.7556	0.4182	1.6941
Proposed	**0.7660**	**0.4432**	**1.8309**

The latter approach provides inferior performance, particularly when compared to the proposed method and the method in [32] indicating the significance of the audio signal for eye-fixation prediction. The proposed method provides improved performance compared to the competing approaches.

C. Discussion

Talmon and Wu provide in [46] a theoretical analysis based on manifold learning studying the kernel product in the continuous limit assuming the existence of $N \to \infty$ data-points and kernel bandwidths approaching zero $\epsilon_v, \epsilon_w \to 0$. They introduced a distance based on the kernel product, which in this limit, is equivalent to a distance obtained using a single modal manifold learning approach applied to the manifold of *hidden* factors that are common to the two modalities. This result, which implies that the kernel product implicitly represents data according to common hidden factors, is empirically supported by Fig. 4 such that, for example, background movements are almost completely attenuated in the videos.

Kernel CCA is more sensitive to interferences as demonstrated in Fig. 4, where we observe interferences that were wrongly detected by kernel CCA as corresponding to the audio. A possible explanation is that Kernel CCA involves the inversion of the kernel matrices, which poses practical limitations on its calculation and often requires the use of a regularization term. Indeed, we found in our experiments that kernel CCA did not converge properly when configured with the same kernel bandwidth as the kernel bandwidth used for the kernel product. Moreover, we have empirically found that using relatively large regularization parameter values did not alleviate the convergence problem. Accordingly, we set the bandwidth to 200 and the regularization parameter to the default value $1e^{-5}$, which led in our experiments to the maximal performance. In this context, we note that improved performance of the kernel product compared to kernel CCA was previously reported by Michaeli et al. in [18] for X-Ray microbeam speech data.

Interestingly, we found that an improved performance of the proposed method is obtained by reducing the weight of the spatial and temporal cues, which are based merely on the video signal. Specifically, the results of the proposed method reported in Table I are obtained by assigning the weights 1, 0.4, 0.4 to the audio-visual correspondence map, the spatial map, and the temporal map, respectively. In contrast, reducing these weights in the method in [32] degraded the performance. Namely, accurate estimation of the correspondence between the audio and the video signals has even a more significant role for eye-fixation prediction than that reported in [32].

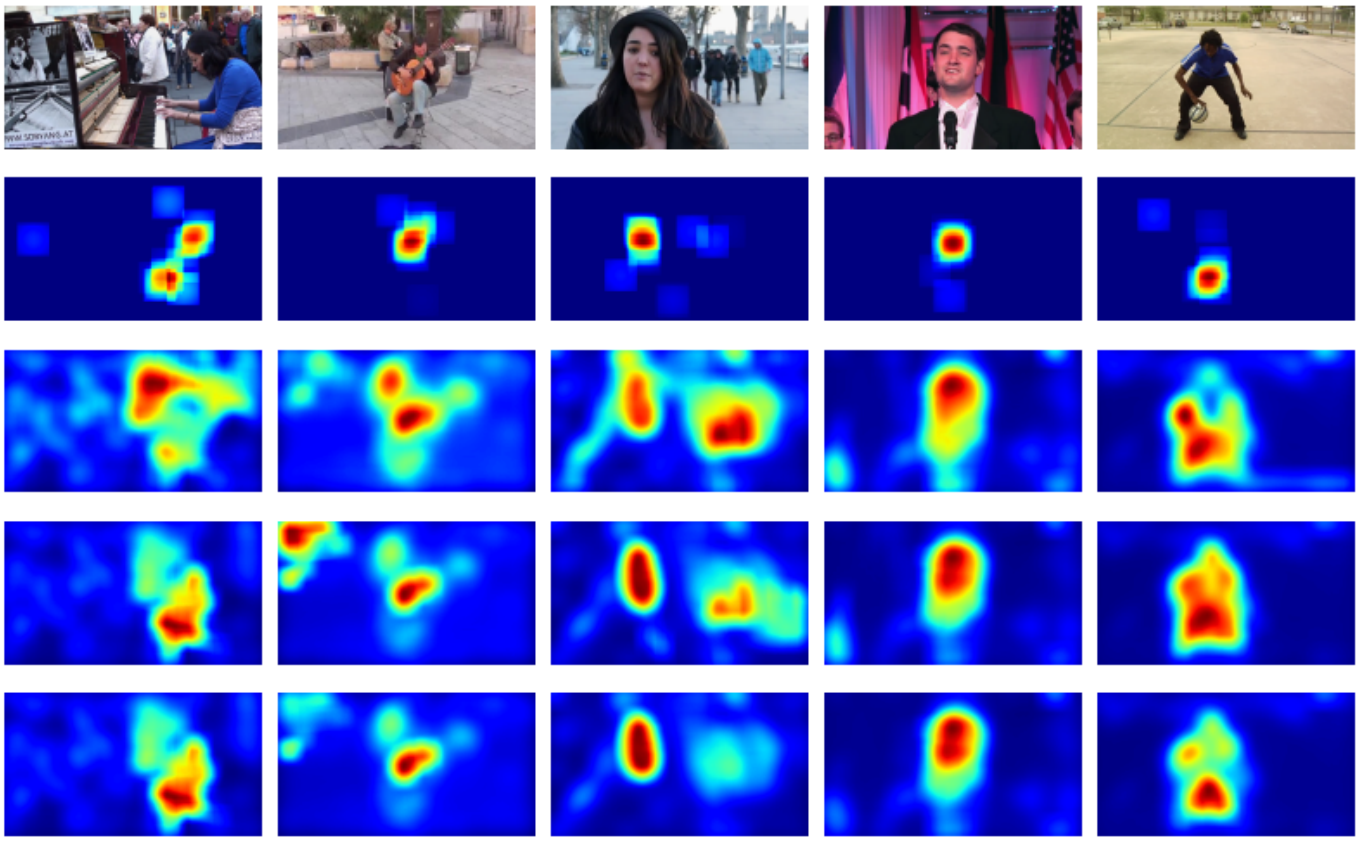
We note in this context that the audio signal contributes to the prediction of eye fixations only when the audio source indeed appears in the video, as we consider in this paper. However, when the audio source is absent from the video, the audiovisual correspondence measure becomes irrelevant and its incorporation may degrade the results. Moreover, the audio source may be present in the video only in part of the time; in such cases, the weights in fusion process between the video-only and the audiovisual measures should be adapted over time. Specifically, one may estimate the existence of an audio source within the video according to the levels of correspondence between the two modalities; then, incorporate the audio-visual correspondence for eye-fixation prediction only if it is above a certain threshold indicating activity of the audio source.

Another important aspect is the influence of the spatial size of the audio source on the locations at which people tend to gaze. Assuming that larger audio sources are more salient, a further improvement in the prediction of eye fixation may be based on weighting the audio-visual correspondence map according to an estimate of the size of the source such that higher weights are assigned to larger audio sources. To further address these aspects of the eye-fixation prediction problem, proper datasets need to be constructed.

In addition, we recall that we use $N = 25$ frames for the construction of the proposed measure of correspondence. The optimal batch size N is set according to a trade-off between the ability to properly capture complex relations between the data-points, i.e., the geometry of the data, and the variability of the signals over time. The derivation in (7), (9) and (11) indicates the contribution of each incoming frame to the measure of correspondence. Setting an adaptive batch size, such that, for example, it can be increased in an online manner by avoiding the subtraction of the last frame in (9) is left for a future research. For example, one may track the variability of the motion vectors over time, and increase the batch size in video regions which are relatively stationary. This may facilitate better learning of the audio-visual geometry improving the accuracy of the correspondence measure.

In this context, for a batch size of $N = 25$, the proposed measure is faster than kernel CCA by almost an order of magnitude as is demonstrated in Fig. 1. This may be significant in online applications such as audio-visual scene analysis, where one would like to detect and separate between several audio-visual sources [62]. Efficient estimation of the correspondence is particularly important since the localization of the audio-visual sources is only a part of a larger online system for source separation.

The speedup of the proposed measure for a batch of $N = 25$ frames, is, however, less significant with respect to the method "single modal update" as can be seen in Fig. 1. With this in mind, we remark that both the proposed measure of correspondence and the corresponding statistical analysis do not make particular assumptions on the type of the modalities. Therefore, we plan to explore the applicability of the proposed



(a) (b) (c) (d) (e)

Fig. 4: Examples of the obtained saliency maps. Each sub-figure corresponds to a different audio-visual recording. From top to bottom: original image, true gaze data (convolved with a Gaussian kernel), a heat map obtained by Min et al. 2016 [32], a heat map obtained by kernel CCA, a heat map obtained by the proposed measure of correspondence.

measure to other modalities in a future research. The optimal number of frames may significantly vary according to both the modalities and the application at hand. Specifically, it depends on the frame rate at which the signals are processed and their variability over time. Consider for example the task of speech enhancement using both a regular and a bone conducting microphone. Multi-modal correspondence may be exploited for the estimation of the spectrum of speech in the presence of transient interferences, which are short term non-speech sounds such as keyboard taps [63]. The frame rate in such a task may be up to 1000 fps as we considered in the single modal setting in [64]. Therefore, we expect the size of the batch to be significantly higher than the one we use here for audio-visual recordings, for which the typical frame rate is $25 - 30$ fps.

VI. Conclusions

We have addressed the problem of measuring correspondence between multi-modal signals in an online setting by proposing a measure based on the trace of the kernel product. We showed how this measure arises in the context of kernel density estimation of data in one modality from the other. In addition, we proposed a statistical model based on the connectivity between data-points showing that the proposed measure is expected to provide high values when signals have a high correspondence in the different modalities. Finally, we proposed an efficient algorithm for online calculation of the proposed measure and demonstrated its improved performance for audio localization in video and for eye-fixation prediction. Future research directions include adaptation over time of the window length used for constructing the measure for each time frame. Namely, the number of frames (the batch size) used for the computation of the kernel may be adapted over time according to dynamical properties of the signal and acoustic conditions. Moreover, the proposed algorithm for online processing allows measuring the contribution to the correspondence measure of each one of the samples (frames). This gives rise to improvement of the proposed measure, e.g., by considering only samples with the highest correspondence levels or by applying time decaying weighting to focus on more recent frames.

Acknowledgment

The authors thank Xiongkuo Min for providing the audiovisual dataset for eye-fixation prediction and the corresponding implementation code. In addition, the authors thank the associate editor and the anonymous reviewers for their constructive comments and useful suggestions.

REFERENCES

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

[2] 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.

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

[4] D. L. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," *Proc. the National Academy of Sciences*, vol. 100, no. 10, pp. 5591–5596, 2003.

[5] R.R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[6] D. Zhou and C. J. C. Burges, "Spectral clustering and transductive learning with multiple views," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 1159–1166.

[7] M. B. Blaschko and C. H. Lampert, "Correlational spectral clustering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.* IEEE, 2008, pp. 1–8.

[8] V. R. De Sa, P. W. Gallagher, J. M. Lewis, and V. L. Malave, "Multi-view kernel construction," *Machine learning*, vol. 79, no. 1-2, pp. 47–71, 2010.

[9] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *Advances in Neural Information Processing Systems*, 2011, pp. 1413–1421.

[10] A. Kumar and H. Daumé, "A co-training approach for multi-view spectral clustering," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 393–400.

[11] Y. Y. Lin, T. L. Liu, and C. S0 Fuh, "Multiple kernel learning for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 6, pp. 1147–1160, 2011.

[12] B. Wang, J. Jiang, W. Wang, Z. H. Zhou, and Z. Tu, "Unsupervised metric fusion by cross diffusion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012*. IEEE, 2012, pp. 2997–3004.

[13] H. C. Huang, Y. Y. Chuang, and C. S. Chen, "Affinity aggregation for spectral clustering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012*. IEEE, 2012, pp. 773–780.

[14] B. Boots and G. Gordon, "Two-manifold problems with applications to nonlinear system identification," *arXiv preprint arXiv:1206.4648*, 2012.

[15] M. M. Bronstein, K. Glashoff, and T. A. Loring, "Making laplacians commute," *arXiv preprint arXiv:1307.6549*, 2013.

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

[17] R. R. Lederman and R. Talmon, "Learning the geometry of common latent variables using alternating-diffusion," *Applied and Computational Harmonic Analysis*, 2015.

[18] T. Michaeli, W. Wang, and K. Livescu, "Nonparametric canonical correlation analysis," in *Proc. International Conference on Machine Learning (ICML)*, 2016.

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

[20] G. Mishne and I. Cohen, "Multiscale anomaly detection using diffusion maps," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 111–123, 2013.

[21] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nyström method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 214–225, 2004.

[22] M. Ding, Z. Tian, and H. Xu, "Adaptive kernel principal component analysis," *Signal Processing*, vol. 90, no. 5, pp. 1542–1553, 2010.

[23] Z. Li, U. Kruger, L. Xie, A. Almansoori, and H. Su, "Adaptive kpca modeling of nonlinear systems," *IEEE Transactions on Signal Processing*, vol. 63, no. 9, pp. 2364–2376, 2015.

[24] L. Xie, Z. Li, J. Zeng, and U. Kruger, "Block adaptive kernel principal component analysis for nonlinear process monitoring," *AIChE Journal*, vol. 62, no. 12, pp. 4334–4345, 2016.

[25] D. R. Perrott, K. Saberi, K. Brown, and T. Z. Strybel, "Auditory psychomotor coordination and visual search performance," *Attention, Perception, & Psychophysics*, vol. 48, no. 3, pp. 214–226, 1990.

[26] J. Vroomen and B. Gelder, "Sound enhances visual perception: cross-modal effects of auditory organization on vision.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 26, no. 5, pp. 1583, 2000.

[27] A. Coutrot and N. Guyader, "Toward the introduction of auditory information in dynamic visual attention models," in *Proc. 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*. IEEE, 2013, pp. 1–4.

[28] G. Song, D. Pellerin, and L. Granjon, "Different types of sounds influence gaze differently in videos," *Journal of Eye Movement Research*, vol. 6, no. 4, 2013.

[29] A. Coutrot and N. Guyader, "How saliency, faces, and sound influence gaze in dynamic social scenes," *Journal of Vision*, vol. 14, no. 8, pp. 5–5, 2014.

[30] X. Min, Gao Z. Zhai, G. and, Hu C., and Wang X., "Sound influences visual attention discriminately in videos," in *Proc. IEEE Int. Workshop on Quality of Multimedia Experience*. IEEE, 2014, pp. 153–158.

[31] X. Min, G. Zhai, Hu C., and Gu K., "Fixation prediction through multimodal analysis," in *Proc. IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2015, pp. 1–4.

[32] X. Min, G. Zhai, K. Gu, and X. Yang, "Fixation prediction through multimodal analysis," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 1, pp. 6, 2016.

[33] H. Izadinia, I. Saleemi, and M. Shah, "Multimodal analysis for identification and segmentation of moving-sounding objects," *IEEE Transactions on Multimedia*, vol. 15, no. 2, pp. 378–390, 2013.

[34] H. Zhang, W. Zhang, W. Liu, X. Xu, and H. Fan, "Multiple kernel visual-auditory representation learning for retrieval," *Multimedia Tools and Applications*, vol. 75, no. 15, pp. 9169–9184, Aug 2016.

[35] E. Kidron, Y. Schechner, and M. Elad, "Pixels that sound," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2005, vol. 1, pp. 88–95.

[36] E. Kidron, Y. Schechner, and M. Elad, "Cross-modal localization via sparsity," *IEEE Transactions on Signal Processing*, vol. 55, no. 4, pp. 1390–1404, 2007.

[37] Zohar Barzelay and Yoav Y Schechner, "Harmony in motion," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR).* IEEE, 2007, pp. 1–8.

[38] G. Monaci, P. Vandergheynst, and F. T. Sommer, "Learning bimodal structure in audio–visual data," *IEEE Transactions on Neural Networks*, vol. 20, no. 12, pp. 1898–1910, 2009.

[39] M. J. Beal, N. Jojic, and H. Attias, "A graphical model for audiovisual object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 828–836, 2003.

[40] E. D'Arca, N. M. Robertson, and J. R. Hopgood, "Robust indoor speaker recognition in a network of audio and video sensors," *Signal Processing*, vol. 129, pp. 137–149, 2016.

[41] R. Cutler and L. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2000, vol. 3, pp. 1589–1592.

[42] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning*, 2013, pp. 1247–1255.

[43] I. Ariav, D. Dov, and I. Cohen, "A deep architecture for audio-visual voice activity detection in the presence of transients," *Signal Processing*, vol. 142, pp. 69–74, 2018.

[44] Z. Lahner, E. Rodola, F. R. Schmidt, M. M. Bronstein, and D. Cremers, "Efficient globally optimal 2d-to-3d deformable shape matching," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2185–2193.

[45] M. Vestner, R. Litman, E. Rodola, A. Bronstein, and D. Cremers, "Product manifold filter: Non-rigid shape correspondence via kernel density estimation in the product space," *arXiv preprint arXiv:1701.00669*, 2017.

[46] R. Talmon and H. Wu, "Latent common manifold learning with alternating diffusion: analysis and applications," *to appear in Applied and Computational Harmonic Analysis*.

[47] A. Gretton, O. Bousquet, A. Smola, and B. Scholkopf, "Measuring statistical dependence with hilbert-schmidt norms," in *ALT*. Springer, 2005, vol. 16, pp. 63–78.

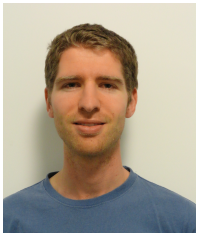
[48] Y. Keller, R. R. Coifman, S. Lafon, and S. W. Zucker, "Audio-visual group recognition using diffusion maps," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 403–413, 2010.

[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[50] *[Online]. Available: https://github.com/lorenzoriano/PyKCCA*.

[51] W. Wang and K. Livescu, "Large-scale approximate kernel canonical correlation analysis," *arXiv preprint arXiv:1511.04773*, 2015.

[52] A. Bruhn, J. Weickert, and C. Schnörr, "Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods," *International Journal of Computer Vision*, vol. 61, no. 3, pp. 211–231, 2005.
[53] A.J. Aubrey, Y.A. Hicks, and J.A. Chambers, "Visual voice activity detection with optical flow," *IET Image Processing*, vol. 4, no. 6, pp. 463–472, 2010.
[54] P. Tiawongsombat, M.H. Jeong, J.S. Yun, B.J You, and S.R. Oh, "Robust visual speakingness detection using bi-level HMM," *Pattern Recognition*, vol. 45, no. 2, pp. 783–793, 2012.
[55] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. 1st International Conference on Music Information Retrieval (ISMIR)*, 2000.
[56] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," 1976.
[57] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1, pp. 23–43, 1998.
[58] J. P. Barker and F. Berthommier, "Evidence of correlation between acoustic and visual features of speech," in *In Proc. of the Int. Congress of Phonetical Sciences*, 1999, pp. 199–202.
[59] C. Xu, C. Xiong, and J. Corso, "Streaming hierarchical video segmentation," *Proc. European Conference on Computer Vision (ECCV)*, pp. 626–639, 2012.
[60] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007, pp. 1–8.
[61] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision Research*, vol. 45, no. 18, pp. 2397–2416, 2005.
[62] D. Dov, R. Talmon, and I. Cohen, "Multimodal kernel method for activity detection of sound sources," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1322–1334, 2017.
[63] D. Dov and I. Cohen, "Voice activity detection in presence of transients using the scattering transform," in *IEEE 28th Convention of Electrical and Electronics Engineers in Israel (IEEEI), 2014*, Dec 2014, pp. 1–5.
[64] A. Hirszhorn, D. Dov, R. Talmon, and I. Cohen, "Transient interference suppression in speech signals based on the OM-LSA algorithm," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4.



David Dov received the B.Sc. (Summa Cum Laude) and M.Sc. (Cum Laude) degrees in electrical engineering from the Technion - Israel Institute of Technology, Haifa, Israel, in 2012 and 2014, respectively. He is currently pursuing the PhD 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 geometric methods for data analysis, multi-sensors signal processing, speech processing, and multimedia.

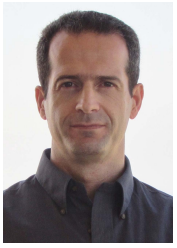
David Dov is the recipient of the IBM PhD Fellowship for 2016-17, the Jacobs Fellowship for 2014, the Excellence in Teaching Award for outstanding teaching assistants in 2013, the Meyer Fellowship, the Cipers Award and the Finzi Award for 2012, the Wilk Award for excellent undergraduate project from the Signal and Image Processing Lab (SIPL), Electrical Engineering Department, Technion for 2012, and Intel Award for excellent undergraduate students for 2009.

Ronen Talmon is an Assistant Professor of electrical engineering at the Technion – Israel Institute of Technology, Haifa, Israel. He 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.

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 2013, he was a Gibbs Assistant Professor at the Mathematics Department, 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), and a coauthor of *Fundamentals of Signal Enhancement and Array Signal Processing* (Wiley-IEEE Press, 2017). 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 Norman Seiden Prize for Academic Excellence, 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 served as Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and IEEE SIGNAL PROCESSING LETTERS, and as a member of the IEEE Speech and Language Processing Technical Committee.