Multi-modal Signal Processing on Manifolds

David Dov

Multi-modal Signal Processing on Manifolds

Research Thesis

Submitted in partial fulfillment of the requirements for

degree Doctor of Philosophy

David Dov

Submitted to the Senate of the Technion—Israel Institute of Technology

Tamuz 5778 Haifa July 2018

Acknowledgment

The Research Thesis Was Done Under The Supervision of Professor Israel Cohen and Professor Ronen Talmon in the Department of Electrical Engineering.

I would like to express my gratitude to my advisers for the supervision, guidance and support throughout this research.

I would to thank Xiongkuo Min for providing the audiovisual dataset for eye-fixation prediction and the corresponding implementation code.

The generous financial support of the Technion, the Jacobs Fellowship, The Israeli Science Foundation (Grants no. 576/16 and 1490/16) and the ISF-NSFC joint research program (grant No. 2514/17) is gratefully acknowledged.

List of papers

- D. Dov, R. Talmon, and I. Cohen, "Sequential Multi-modal Correspondence With Alternating Diffusion Kernels", IEEE Transactions on Signal Processing, vol. 66, no. 12, pp. 3100-3111, June 2018.
- D. Dov, R. Talmon, and I. Cohen, "Multi-modal Kernel Method for Acoustic Scene Analysis", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 6, pp. 1322–1334, June 2017.
- 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.
- D. Dov, R. Talmon, and I. Cohen, "Kernel method for voice activity detection in the presence of transients," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 12, pp. 2313–2326, Dec 2016.
- D. Dov, R. Talmon, and I. Cohen, "Kernel Method for Speech Source Activity Detection in Multi-modal Signals", in Proc. IEEE International Conference on the Science of Electrical Engineering (ICSEE 2016)

Contents

List of Figures

List of Tables

Abstract

Multi-modal signals, i.e., signals measured by multiple sensors of different types, often have different characteristics across the modalities, e.g., different dynamics, dimensions, and value range. Accordingly, various sources of data are expressed differently across the modalities such that part of them are common to the different modalities and others appear only in specific modalities. For example, a speech signal measured by a microphone and a video camera is considered a common source, while speech from other speakers is audio-specific. These unique characteristics are appealing in various applications such as the analysis of audio-visual scenes, but also raises fundamental questions related to the joint analysis of the multi-modal signals. The questions which we address in this thesis are how to obtain a representation of the signal according to the common source, while reducing the effect of modality-specific sources considered interferences; how to process data available in the different modalities only in certain time intervals; and how to measure to what extent the data in the different modalities correspond ("correlate") to each other.

We address these questions from manifold learning perspective by the design of kernel-based geometric methods. Classical kernel-based methods are typically designed for analyzing data measured in a single sensor by learning geometric structures of high dimensional data. These methods provide low dimensional representations of the data via the eigenvalue decomposition of affinity kernels capturing local relations (affinities) between samples of the

signal.

In this thesis, we address the problem of data fusion via the combination of affinity kernels constructed separately for each modality. We consider a particular combination of the kernels via the product function previously shown to provide a representation of data according to the common source. We introduce a new graph-theoretic interpretation to this approach relating the kernels to their corresponding single and multi-modal graphs. We analyze the relations between the connectivities of the graphs, and, based on this analysis, we further improve the fusion process by an improved method for the construction of the affinity kernels. Then, we extend the context of the fusion problem to a setting, where the data in the different modalities is only partially available. We show how the proposed fusion approach can be extended to this setting allowing to obtain a joint representation of signals, according to the common source, even when the multi-modal signals are available only in certain time intervals. Finally, we address the question to what extent signals from different sensors correspond to each other, i.e, contain similar content. By revisiting the graph-theoretic analysis of the product of kernels, we devise a measure for multi-modal correspondence and show how it can be efficiently updated in an online setting. We demonstrate the proposed approaches for data fusion and for measuring multimodal correspondences for various applications related to the analysis of complex audio-visual sound scenes. In particular, we report improved performance for different variants of the task of sound sources activity detection and audio localization in video.

Notations

B	Binomial distribution
C	Constant related to the selection of the kernel bandwidth
$\mathbf{C}$	Covariance matrix
$\mathcal{C}$	Discrete set of values on a linear grid
$d(\cdot,\cdot)$	Diffusion distance
$d_v(\cdot)$	Row normalization factor of the first modality
$\mathbf{D}$	Row normalizing matrix
$\mathbb{E}(\cdot)$	Expected value
$f(\cdot)$	Unknown (possibly non-linear) mapping
$g(\cdot)$	Unknown (possibly non-linear) mapping
$h(\cdot)$	Unknown (possibly non-linear) mapping
$\mathbf{h}_n$	Vector describing propagation of diffusion in the nth step
$\mathbb{1}$	Indicator
$\mathbf{K}$	Affinity kernel
L	Dimension of data
$\mathbf{M}$	Row normalized affinity kernel
M^v	Number of sources in the first modality
N	Number of frames
p	Probability
$\mathbf{P}$	Assignment matrix
q	Probability that a point is disconnected
S_v	Average number of connections in a single-modal graph

S_i	The ith source of data
S^d	Desired source
Tr	Trace
v, w	General notation of the first and the second modality, respectively
$(\mathbf{v}_n, \mathbf{w}_n)$	pair of multi-modal data points in time index n
δ	Average number of connections in a multi-modal graph
ϵ	Kernel bandwidth
$\mathcal{H}_0$	Hypothesis that a source is not active
$\mathcal{H}_1$	Hypothesis that a source is active
λ	Eigenvalue
ν	Eigenvector
σ	Singular value
τ	Threshold value
$\mathbf{\Phi}$	Diffusion maps matrix
$\boldsymbol{\psi}_n$	Diffusion maps of frame n
$\circ$	Hadamard product
$\|\cdot\|$	Euclidean normalization
$(\cdot)^T$	Transpose

Unique notations in Chapter 5

a	General notation of audio signals
$(\mathbf{a}_n, \mathbf{v}_n)$	Pair of audio-visual data points in time index n
L	Number of frames in a reference set
Q	Number of interfering video sources
R	Number of interfering audio sources
s_i	The ith source of data
$\tilde{s}$	Source of interest
v	General notation of video signals

θ	Quantity computed over the reference set
ϕ	Eigenvectors

Unique notations in Chapter 6

f	Density function
$\tilde{()}$	Updated quantity
$\mathbb{I}$	Indicator
$\mathbb{J}$	Measure for the connectivity of the normalized kernel

Unique notations in Chapter 7

d	Dimension of the vector of the latent variables
J	Dimension of the obtained representation
$\mathbf{J}$	Jacobian matrix
t	General notation for transient interferences
x	General notation for speech
$\mathbf{y}_n$	Feature representation of an audio signal in frame n
r	Constant factor encoding the dominance of transients
R	Quantity defining the size of the temporal window
$\boldsymbol{\theta}$	Vector of latent variables
$\boldsymbol{\Lambda}$	Quantity related to the Jacobian matrix
$\boldsymbol{\mu}$	Mean vector
σ^2	Variance
$\boldsymbol{\phi}$	Eigenvector
$\boldsymbol{\psi}$	Vector of normalized latent variables

Abbreviations

AUC	Area Under the Curve
CC	Correlation Coefficient
CCA	Canonical Correlation Analysis
DCT	Discrete Cosine Transform
HSIC	Hilbert-Schmidt independence criterion
IID	Identically and Independently Distributed
MFCC	Mel-Frequency Cepstral Coefficients
MMSE	Minimum Mean Square Error
NSS	Normalized Scanpath Saliency
PCA	Principal Component Analysis
ROC	Receiver Operating Characteristic
sAUC	shuffled Area Under the Curve
SIR	Source of interest to Interferences Ratio
SNR	Signal to Noise Ratio
STFT	Short-Time Fourier Transform
SVM	Support Vector Machines
VAD	Voice Activity Detector

Chapter 1

Introduction

Multi-modal signal processing is a field of research focusing on the joint analysis of signals measured by multiple sensors of different types. In recent years, this field has gained an increased attention from both the signal processing and the data analysis communities due to an extensive use of multiple sensors in various fields such as medicine, entertainment, defense and autonomous driving. For example, multiple microphones and video cameras are regularly integrated in laptops and smartphones, as well as in autonomous vehicles along with other sensors such as GPS and different types of accelerometers. Other examples are Electroencephalography (EEG), Electrooculography (EOG) and Electromyography (EMG) sensors used in medicine for example to assess sleep quality of patients.

Due to the differences in the types of the sensors, multi-modal signals often have different characteristics, e.g., different dynamics, dimensions and amplitude ranges. They capture rich and valuable information on the measured phenomenon, which may be joint to the different modalities or complementary and specific to a particular sensor. While these unique properties make multi-modal signals highly appealing in various applications, there are fundamental challenges and open questions related to their processing such as how to fuse data from the different modalities; how to process data which is

available in the different modalities only in certain time intervals; and how to measure the extent of correspondence between the different modalities. This research was directed at addressing these open questions by the analysis and design of kernel-based geometric methods.

1.1 Manifold learning for multi-modal signal processing

Often referred to as manifold learning, classical kernel-based geometric methods such as those presented in [13, 18, 34, 45, 110] are typically designed for the analysis of single modal data. They are based on the construction of an affinity kernel, which captures the geometry of the data, i.e., local relations (affinities) between the data-points. Via the eigendecomposition of the kernel, these methods provide low dimensional representations of the data, preserving the local affinities. Since data such as images and speech signals admit certain geometric structures, manifold learning methods are successfully applied to a variety of applications such as anomaly and target detection and speech enhancement [96, 97, 131, 133]. In the remainder of this chapter, we describe the questions explored in this research for developing kernel based geometric methods for multi-modal signal processing and indicate the scientific gaps that motivated our work.

A fundamental question, related to the analysis of multi-modal signals, is how to obtain a joint representation of signals via the fusion of data acquired in different modalities. We are particularly interested in a scenario where the signals are corrupted with structured interferences and assume that each type of interferences appears only in one of the sensors. We address this question from the manifold learning perspective via the construction of affinity kernels separately for each modality and study how to combine these kernels to obtain a joint representation of the signals.

In particular, we focus on the combination of kernels via kernel multipli-

cation. This fusion approach was shown in [79] to provide a representation of signals according to factors which are common to the two modalities. Yet, the analysis in [79], which was done in the continuous domain, did not shed light on important aspects of the fusion process such as how to construct the affinity kernel and in particular how to choose the kernel bandwidths which are important parameters controlling the scale at which the geometry of the data is learned. To address these questions, we propose a new analysis of this fusion approach based on graph theory by associating the single and the multi-modal kernels to graphs whose vertices are the data-points. Our analysis relates the connectivity of the graphs to the kernel bandwidths allowing us to introduce an algorithm for their proper selection, which further improves the fusion process between the modalities.

Another fundamental question in manifold learning for signal processing is how to extend the kernel methods to new samples. Typically, the geometric structure of data is learned in an offline manner using a batch of samples acquired in advance. Then, extension approaches such as the Nyström method [55] are used to calculate the representation of new incoming samples. We address the question of out of sample extension in the multi-modal setting considering a scenario where the data from the different sensors are available only in certain time intervals. We show how to obtain a joint representation from a short time interval containing data from all of the sensors, and then, extend the joint representation to new samples even if they are available only from one sensor.

As an application of the proposed fusion approach, we consider the analysis of audio-visual sound scenes comprising multiple sound sources. We consider sound sources of different types including multiple speakers, various environmental noises, and transients, which are short-term interferences such as keyboard taps. We assume that the video camera is pointed to a particular source to which we refer as the source of interest, while all other sources are considered interferences. The goal is to detect the activity of the common

source while neglecting the other sources. Having the video camera pointed at the face of a speaker in the presence of background noise is a special case of the task we consider, which is termed voice activity detection and is a widely studied problem in the speech processing community [108,123,136]. However, here we consider more general scenarios such as the presence of interfering speakers, which is difficult to distinguish from the speech of interest since they often share similar characteristics. We show that the proposed fusion approach indeed provides a representation, in which the interfering sources are attenuated, and thus, allowing us to introduce an activity measure, which accurately detects the source of interest.

Then, we consider the question to what extent signals from different sensors correspond to each other, i.e, contain similar content. We show how such a measure of correspondence arises from the theoretical analysis of the graphs related to the fusion process via the product of kernels. In an online setting, classical kernel methods are often based on eigenvalue decomposition, which due to computational constraints cannot be recomputed for each new incoming frame. Classical out of sample extension methods such as the Nyström method [55] only approximate the representation of a new sample via interpolation from a training set. The proposed measure of correspondence does not require eigendecomposition so that it is particularly suitable for an online setting, and we further show how to efficiently update it over time. We demonstrate the use of the proposed measure in two related applications. First is audio localization in video, which we formulate as a correspondence task by dividing the video into separate spatial regions and measure the correspondence of each region to the audio. The second application is eye fixation prediction, in which the goal is to predict regions at which people gaze while watching videos. This application is highly related to the problem of audio localization in video since studies in psychology [35, 36, 91–93, 103, 125, 144] imply that people tend to gaze at the audio sources while watching videos. We show that high correspondence levels are indeed obtained for regions in

which the audio source is located.

In the final part of this research, we consider the single sensor setting and address two fundamental questions that should be addressed in order to obtain a good representation of signals using kernels based methods. First is the question of how to capture the dynamics of the signal, i.e., temporal variations over time. Classical kernel methods measure affinities between signals while treating the samples as data points and neglecting the temporal relations between them. The second question is how to choose a proper metric to measure the affinities between the samples, which is crucial for obtaining a good representation. We address these questions in the context of obtaining a representation of signals contaminated by interferences, in which the signal and the interferences are properly distinguished. From a kernel based geometric standpoint, the key element in obtaining such a representation is finding a metric (and consequently an affinity kernel) that appropriately distinguishes between the signal and the interferences. We consider a challenging example of the representation of speech signals contaminated by short-term interferences such as keyboard taps, which through the Euclidean distance wrongly appear similar to each other, thereby resulting in a poor distinction between them. To address this problem, we propose to use a metric based on the statistics of the signal in short temporal windows, which we assume to be different between the signal and the interferences. We justify the use of this metric by introducing a model of independent latent variables related to the signal and the interferences. These variables may be related to physical constraints controlling the generation of the signal, and, as such, they accurately represent the content of the signal. We show that the proposed metric approximate distances between samples of the noisy signal according to the latent variables and thus allowing to properly represent the noisy signals according to its content.

1.2 Main Contributions

The main contributions of this thesis are as follows. We address the problem of sensor fusion by incorporating data via the product of kernels, constructed separately for each modality [48]. We propose a new analysis to this fusion approach based on graph theory, where the single and the multi-modal kernels are associated with graphs whose vertices are the data-points. We analyze the relation between the connectivity of the single and the multi-modal graphs and link between the connectivity of the graphs and the kernel bandwidths. The selection of these parameters is particularly important since learning geometry at too fine scales may lead to a noisy and an unstable representation while too rough scales cannot properly capture the geometry of the signals. While the selection of the kernel bandwidth in the multi-modal case was not addressed in the literature to the best of our knowledge, we introduce an algorithm for its proper selection, which further improves the fusion process between the modalities.

Then we extend the scope of the fusion problem to an online setting considering a practical scenario where the data from the different sensors are available only in certain time intervals [50]. Interestingly, we show that a joint representation of the data may be successfully learned even from a short time interval containing data from all of the sensors, and, in turn, extended to new samples from a single modality. As an application of the proposed fusion approach, we consider the task of sound source activity detection in audio-visual recordings [48–50]. We show that the proposed fusion approach indeed provides a joint audio-visual representation, where the effect of the interfering sources is attenuated, and thus it allows for accurate activity detection of the source of interest.

Then, we propose a measure of correspondence between multi-modal signals based on the trace of the kernel product and show how variants of this measure are motivated both by the theoretical analysis of the corresponding graphs and by the interpretation of the kernel product as an estimator of one

modality based on the other [51]. We further show how to efficiently update the proposed measure over time in an online setting. We demonstrate the use of the proposed measure for the tasks of audio localization in video and eye fixation prediction showing that the proposed measure successfully indicates the salient regions and outperforming competing methods.

Finally, we consider the problem of obtaining a representation of signals contaminated by structured interferences, in which the interferences are properly distinguished from the signals and attenuated [52]. We propose to use a metric based on the statistics of the signal in short temporal windows, and justify it by modeling the signal and the interferences by their latent generating variables. We show that the Euclidean distance between the latent variables is approximated by the proposed metric. Moreover, assuming that the interferences have fast variation rates over time, we further show that the proposed metric reduces their effect on the obtained representation. With this improved metric, we address the task of voice activity detection in the presence of interferences such as keyboard taps. By incorporating the proposed metric into a kernel-based manifold learning method, we devise a measure of voice activity, which outperforms competing detectors for different types of interferences.

1.3 Thesis Structure

The remainder of the thesis is organized as follows. In Chapter 2 we provide a brief theoretical background of manifold learning and its use for signal processing. In Chapter 3, we address the fusion of multi-modal signals for obtaining a joint representation via a product of affinity kernels constructed separately for each modality. We analyze this fusion approach from a graph-theoretic perspective and further improve it by proposing an algorithm for the selection of the kernel bandwidth. With the improved representation, we address the problem of audio-visual voice activity detection, which is a spe-

cial case of the problem of activity detection of sound sources, in which the source whose activity is detected is speech and the other sources are background noise and transient interferences. In Chapter 4 we further extend our experiments to a setting where the interfering sources are speech from other speakers, which is even more challenging setting since the signal and interferences have similar statistical and temporal characteristics. In Chapter 5, we consider an online setting in which the signals in the different modalities are available only in certain time intervals. We propose to use short intervals in which the signal is available across modalities for obtaining the joint representation and then show how to extend it to new samples from single modalities.

Then, in Chapter 6, we address the question of how to measure the correspondence between multi-modal signals in an online setting. We propose a measure of multi-modal correspondence based on the trace of the product of kernels. We show how the proposed measure is related to kernel density estimation and justify the measure from a graph-theoretic point of view. Then, we show how to efficiently calculate it in an online manner. We demonstrate the use of the proposed measure in applications of audio source localization in videos and eye-fixation prediction and show that it outperforms competing methods.

In Chapter 7 we address the question of how to obtain a representation of signals in the presence of interferences in a single modal setting. We consider the example of speech signals contaminated by short-term interferences focusing on the design of a metric which properly distinguishes between the signals and the interferences by exploiting differences between their short-term statistics. We analyze this metric using a model of underlying factors controlling the generation of the signal. By plugging the proposed metric into a kernel based geometric method we obtain a representation with a good distinction between the signal and the interferences, which leads to improved performance in the application of single modal voice activity detection. We

conclude our research and discuss future research directions in Chapter 8.

Chapter 2

Scientific Background

This research addresses the design and analysis of kernel-based geometric methods for multi-modal signal processing. Here, we bring a brief scientific background of kernel-based geometry learning for signal processing. We focus on diffusion maps [34], which is a classical method for geometry learning, and on its extension to the multi-modal case.

2.1 Kernel-Based Geometry Learning for Signal Processing

Classical kernel-based geometric methods, e.g., those presented in [13, 18, 34, 45, 110], represent a class of nonlinear dimensionality reduction methods designed for analyzing data measured by a single sensor. Consider a sequence of N samples of a signal $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_N$, where $\mathbf{v}_n \in \mathbb{R}^{L_v}$ is the nth sample, and v denotes the sensor. By learning geometric structures of high dimensional data, these methods typically provide low dimensional representations of the data, $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, ..., \boldsymbol{\psi}_N$, where $\boldsymbol{\psi}_n \in \mathbb{R}^L$ and $L \ll L_v$, via the eigenvalue decomposition of an affinity kernel. The low dimensional representations preserve the geometry of the signals, i.e., local affinities between their samples, and

they are successfully used in a wide range of applications such as anomaly and target detection and speech enhancement [96, 97, 131, 133]. For the sake of completeness, we briefly describe a certain method termed "diffusion maps" which we exploit in our research. Let $K_v(n,m)$ be an affinity kernel between frames $\mathbf{v}_n$ and $\mathbf{v}_m$, typically given by:

$$K_v(n,m) = e^{-\frac{\|\mathbf{v}_n - \mathbf{v}_m\|^2}{\epsilon_v}}, \tag{2.1}$$

where ϵ_v is the kernel bandwidth - a scaling parameter chosen, e.g., according to [67]. Short distances between frame $\mathbf{v}_n$ and frame $\mathbf{v}_m$ provide high values of the kernel, whereas distances much greater than the scaling parameter ϵ_v are negligible. We note that the selection of the affinity kernel and particularly the kernel bandwidth are of key importance for properly learning the geometric structure of the signal. In Chapter 3 we discuss in detail why and how we deviate from the typical choice of the kernel bandwidth in the multi-modal setting and in Chapter 7 we analyze a specially designed metric, which allows for improved distinction between signals and interferences.

Using the kernel in (2.1), an affinity matrix $\mathbf{K}_v \in \mathbb{R}^{N\times N}$ is constructed such that its (n,m)th entry, represents the affinity between frame $\mathbf{v}_n$ and frame $\mathbf{v}_m$ and is given by $K(\mathbf{v}_n, \mathbf{v}_m)$. The affinity matrix $\mathbf{K}_v$ defines a weighted symmetric graph such that the frames $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_N$ are the nodes of the graph and the edge between frame $\mathbf{v}_n$ and frame $\mathbf{v}_m$ is given by $K_v(n,m)$. A Markov chain on the graph is defined by normalizing the kernel [34]:

$$\mathbf{M}_v = \mathbf{D}_v^{-1}\mathbf{K}_v, \tag{2.2}$$

where $\mathbf{D}_v \in \mathbb{R}^{N\times N}$ is a diagonal matrix with $D_v(m,m) = \sum_n K_v(m,n)$. Namely, $\mathbf{M}_v \in \mathbb{R}^{N\times N}$ is a row stochastic Markov matrix whose rows sum to one. The normalized affinity kernel $\mathbf{M}_v$ captures the relations between the samples of the signals in the form of probabilities of transitioning from one sample (node) to another.

Finally, an eigenvalue decomposition is applied to $\mathbf{M}_{\mathbf{v}}$, yielding eigenvalues $\{\lambda_l\}$, which are sorted in descending order, and corresponding eigenvectors $\{\boldsymbol{\nu}_l\}$. The eigenvalues of $\mathbf{M}$ are in the range of $0 \div 1$ due to the row normalization [34]. Moreover, $\lambda_0 = 1$ and its associated eigenvector $\boldsymbol{\nu}_0$ is an all-ones vector. This constant eigenvector is ignored since it does not contain any information [77].

The first L largest eigenvalues (excluding the trivial) and the corresponding L eigenvectors of $\mathbf{M}$ are used for the parametrization of the data points. Let $\boldsymbol{\Phi} \in \mathbb{R}^{N \times L}$ be a matrix whose columns consist of the eigenvectors and the eigenvalues of $\mathbf{M}_v$:

$$\boldsymbol{\Phi} \equiv (\lambda_1 \boldsymbol{\nu}_1, \lambda_2 \boldsymbol{\nu}_2, ..., \lambda_L \boldsymbol{\nu}_L)\,. \tag{2.3}$$

From (2.3), the diffusion mapping of sample $\mathbf{v}_n$, which we denote by $\boldsymbol{\theta}_n$ is given by the nth row of the matrix $\boldsymbol{\Phi}$:

$$\boldsymbol{\psi}_n = (\boldsymbol{\Phi}_{n,1}, \boldsymbol{\Phi}_{n,2}, ..., \boldsymbol{\Phi}_{n,L})\,. \tag{2.4}$$

The embeddedings (mappings) $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, ..., \boldsymbol{\psi}_N$ corresponding to the samples $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_N$ are viewed as the new L dimensional representation of the signal. Assuming that there exists a low dimension geometric structure of the signals, the eigenvalues decays fast, and L (2.3) may be set to a small value providing significant reduction of dimensionality.

The use of diffusion maps is motivated through the diffusion distance. For simplicity, we describe here an unnormalized spacial case of the more general diffusion distance, presented in [34]. Let $d\,(n, m)$ be the diffusion distance between frame n and frame m, given by [79]:

$$d\,(n, m) = \sqrt{\sum_{k=1}^{N} \left(M_v(n, k) - M_v(m, k)\right)^2}. \tag{2.5}$$

According to (2.5), the distance between frame n and frame m is roughly given by a collection of transition probabilities in one step between the frames. The distance between each pair of frames is based on transition probabilities to all other frames in the set respecting the geometry of the data, and as a result, it is considered robust to noise [34]. The use of diffusion maps is motivated in [34] by showing that the l_2 distance in the diffusion maps domain is equivalent to the diffusion distance for $L = N$:

$$|\boldsymbol{\psi}_n - \boldsymbol{\psi}_m| = d(n, m).$$

According to (2.1)-2.4, diffusion maps are constructed in a batch manner assuming that N samples of the signal are available in advance. In an online setting, diffusion maps are typically extended to new incoming samples using interpolation techniques such as the Nyström method [55]. Let $\mathbf{v}_{N+1}$ be a new incoming sample; its representation in the diffusion maps domain is given by extending the eigenvectors of $\mathbf{M}_v$ in (2.2) to the new sample:

$$\nu_l(N+1) = \frac{1}{\lambda_l} \sum_{n=1}^{N} M_v(N+1, n) \nu_l(n),$$

where $\nu_l(N+1)$ may be viewed as a weighted interpolation of:

$$\{\nu_l(1), \nu_l(2), ..., \nu_l(N)\},$$

which are the entries of the eigenvector $\boldsymbol{\nu}_l$ constructed from the samples $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_N$. The weights of the interpolation are given by the affinities $M_v(N+1, n)$, $n = 1, 2, ..., N$ between the new sample $N+1$ and the batch samples.

2.2 Kernel-Based Multi-modal Sensor Fusion

The potential of improving the robustness of the obtained representations to interferences by fusing data captured in multiple modalities has recently motivated researchers to extend kernel-based geometric methods to the multi-modal case [19, 20, 42, 61, 67, 72, 73, 77, 79, 81–83, 90, 145, 153]. These studies share similar ideas; broadly, they are based on constructing separate affinity kernels for each modality, and then fusing the data by a certain combination of the affinity kernels or their eigenvectors, e.g., by the product between affinity kernels or by their weighted sums. A method of particular interest in this study is the *alternating diffusion maps*, presented in [79] for data fusion. The alternating diffusion maps method is designed to reveal the geometric structure of the data, which is common to two modalities, while ignoring the interferences, which are captured only in one of the modalities. In the following, we shortly describe the construction of alternating diffusion maps. Consider a dataset of N samples captured in two different modalities (sensors) given by:

$$(\mathbf{v}_1, \mathbf{w}_1), (\mathbf{v}_2, \mathbf{w}_2), ..., (\mathbf{v}_N, \mathbf{w}_N), \tag{2.6}$$

where, similarly to $\mathbf{v}_n$, $\mathbf{w}_n \in \mathbb{R}^{L_w}$ is the nth sample of the signal measured by the second sensor. An example we addressed under this setup is an audio-visual recording of a speaker, where $\mathbf{v}_n$ is the nth time frame of the signal captured in a microphone and $\mathbf{w}_n$ is the corresponding video frame of the mouth region of the speaker.

Similarly to $\mathbf{K}_v \in \mathbb{R}^{N\times N}$ in (2.1) and $\mathbf{M}_v \in \mathbb{R}^{N\times N}$ in (2.2), let $\mathbf{K}_w \in \mathbb{R}^{N\times N}$ and $\mathbf{M}_w \in \mathbb{R}^{N\times N}$ be the affinity matrix and the corresponding row stochastic matrix (2.1) of the second modality, respectively. The modalities are fused by the product of the row stochastic matrices [79]:

$$\mathbf{M} = \mathbf{M}_v \mathbf{M}_w, \tag{2.7}$$

where $\mathbf{M}$ is the unified kernel, which is also row stochastic, integrating the relations between the data points over the two modalities [79, 135]. The continuous counterparts of the matrices $\mathbf{M}_v$, $\mathbf{M}_w$ and $\mathbf{M}$ are typically considered in the literature as diffusion operators [34]. The authors in [79] introduced an analysis showing that the unified kernel $\mathbf{M}$ in 2.7 is equivalent to an alternating diffusion operator consisting of two diffusion steps on the two modalities. They showed that this alternating diffusion attenuates the modality-specific interferences and, thus, provides a representation of the signals according to their common factors. This analysis, however, does not shed light on the question of how to choose the affinity kernels and what effect the amplitudes of the interferences have on their selection. In Chapter 3 we address these questions by introducing a graph-theoretic analysis of the alternating diffusion maps method.

2.3 The Mahalanobis Distance

Both as reported in previous studies [132] and as we observed throughout numerous experiments, the selection of the affinity kernel plays a crucial role in the ability of kernel methods to provide useful representations of signals. In Chapter 7, we consider the representation of speech signals in the presence of structured interferences, where the goal is to properly distinguish between them. We show that this goal cannot be achieved using a standard affinity kernel $\mathbf{K}_v$ (2.1) since it is based on the Euclidean distance, through which speech and interferences wrongly appear similar. To overcome this problem, we propose in Chapter 7 to use a modified version of the Mahalanobis distance, rather than the Euclidean distance (2.1). Denoted by $\|\cdot\|_M^2$, the modified Mahalanobis distance is given by:

$$\|\mathbf{v}_n - \mathbf{v}_m\|_M^2 \triangleq \frac{1}{2} \left(\mathbf{v}_n - \mathbf{v}_m\right)^T \left(\mathbf{C}_n^{-1} + \mathbf{C}_m^{-1}\right) \left(\mathbf{v}_n - \mathbf{v}_m\right),$$

where $\mathbf{C}_n \in \mathbb{R}^{L\times L}$ and $\mathbf{C}_m \in \mathbb{R}^{L\times L}$ are the covariance matrices of $\mathbf{v}_n$ and $\mathbf{v}_m$, respectively. This version of the Mahalanobis distance was first introduced and studied in [119] in the context of the problem of non-linear independence component analysis. The authors showed that if the observable signal $\mathbf{v}_n$ is generated by independent latent stochastic dynamical processes, the Mahalanobis distance approximates distances between samples of the signal in the latent domain. This property is desirable in the context of distinguishing between speech and transients since they are independent of each other so that they can be represented by independent components. However, such processes cannot properly model speech or transients since the state of the processes is assumed to slowly evolve in time, while, for example, transients are typically fast varying and their activity can abruptly transition between presence and absence. In Chapter 7, we assume a (different) statistical model, in which speech and transients are controlled by latent independent variables and show that the Mahalanobis distance approximates a weighted Euclidean distance between the variables allowing to properly represent the signal according to its content. Moreover, similarly to [53], we assume different variation rates of the different components; specifically, we assume that the transients, due to their abrupt nature, vary faster than speech over time, and show that the Mahalanobis distance reduces their effect on the obtained representation.

Chapter 3

Kernel-based Sensor Fusion with Application to Audio-Visual Voice Activity Detection

In this chapter, we address the problem of multiple view data fusion in the presence of noise and interferences. Recent studies have approached this problem using kernel methods, by relying particularly on a product of kernels constructed separately for each view. From a graph theory point of view, we analyze this fusion approach in a discrete setting. More specifically, based on a statistical model for the connectivity between data points, we propose an algorithm for the selection of the kernel bandwidth, a parameter, which, as we show, has important implications on the robustness of this fusion approach to interferences. Then, we consider the fusion of audio-visual speech signals measured by a single microphone and by a video camera pointed to the face of the speaker. Specifically, we address the task of voice activity detection, i.e., the detection of speech and non-speech segments, in the presence of structured interferences such as keyboard taps and office noise. We propose an algorithm for voice activity detection based on the audio-visual signal. Simulation results show that the proposed algorithm outperforms competing

fusion and voice activity detection approaches. In addition, we demonstrate that a proper selection of the kernel bandwidth indeed leads to improved performance.

3.1 Introduction

Multiple view data fusion is the process of obtaining a unified representation of data captured in multiple measurement systems of different types. Data fusion has recently attracted a growing interest in the signal processing and data analysis communities due to an extensive use of multiple sensors in everyday devices such as computers and smartphones. Often, data measured in multiple views is contaminated with noises and interferences which are view specific, and fusing the views may allow for obtaining representations of the data, which are robust to the interferences. A challenging example which we consider in the current work is the fusion of audio and visual recordings of a speaker. While each view (audio or video) possibly consists of view-specific interferences (e.g., acoustic noises or face movements), their fusion may give rise to a robust representation of the speech.

In this chapter, we use a kernel based geometric approach to address the problem of multiple view data fusion. Classical methods, e.g., those presented in [13, 18, 34, 45, 110], represent a class of non-linear dimensionality reduction methods designed for data measured in a single view. By learning geometric structures of high dimensional data, these methods provide low dimensional representations of the data via eigenvalue decomposition of an affinity kernel. The low dimensional representations preserve the geometry of the data, i.e., local affinities between data points, and they are successfully used in a wide range of applications such as anomaly and target detection and speech enhancement [96, 97, 131, 133]. However, when the data is corrupted by structured interferences, the kernel methods learn the structure of the interferences along with the structure of the data. Therefore, the obtained

low dimensional representation retains the relations between the data and the interferences, and, as a result, the kernel methods have a limited robustness to the interferences.

The potential of improving the robustness of the obtained representations to interferences by fusing data captured in multiple views, has recently motivated researchers extending kernel-based geometric methods to the multiple views case [19–21, 42, 61, 72, 73, 79, 81, 83, 90, 145, 153]. Among these studies, we mention the studies presented in [42, 79, 83, 90, 145] sharing similar ideas of constructing separate affinity kernels for each view, and fusing the data by the product between the affinity kernels. A method of particular interest in this work was presented in [79], where special emphasis is given to the robustness of the fusion process to interferences. The authors presented a data fusion method termed alternating diffusion maps, which is based on fusing the views by multiplying between affinity kernels interpreted as employing separate diffusion processes on each view in an alternating manner. By analyzing the method in the continuous setting, it is shown that the interruptions of a certain view are attenuated by the diffusion steps of the other views.

The ability of kernel methods to properly learn the geometric structure of the data is highly dependent on the selection of the kernel bandwidth, which also has important implications on the robustness of the methods to interferences. The kernel bandwidth, also called the scale parameter, defines a local neighborhood such that all data points within the neighborhood are considered similar, i.e., close to each other. It has an intuitive interpretation by viewing the kernel methods from the graph theory point of view, which we adopt throughout this chapter. The affinity kernel defines a graph whose nodes are the data points and the edges are given by the affinities between the data points. Accordingly, all data points located within a local neighborhood defined by the kernel bandwidth are considered connected on the graph. In the single view case, the kernel bandwidth is chosen according to

a trade-off. On the one hand, it has to be large enough keeping the graph connected, which is a necessary condition for learning the geometry of the data [34, 77, 143, 151]. On the other hand, the kernel bandwidth has to be as small as possible so that data and interferences will not share the same local neighborhoods [67]. In the multiple view case, the selection of the kernel bandwidth is not addressed in the literature, and the kernel bandwidths are naively chosen in previous studies as if the data is measured in a single view.

As an application for multiple view data fusion, we consider in this chapter the problem of audio-visual voice activity detection, in which the goal is to detect time intervals of active speech from audio-visual recordings. A common practice in the analysis of speech measured by a video camera is the design of features modeling the shape and the movement of the mouth. Examples of such features are the height and the width of the mouth [121, 122], key-points at the mouth region [9, 84, 100], intensity levels of the mouth region [117], and motion vectors [10, 137]. Two main approaches that exist in the literature for the incorporation of the audio and video signals, are called early and late fusion [8, 136]. In early fusion, features, constructed from the audio and video signals, are concatenated into a single feature vector and are viewed as data obtained in a single view [6]. In late fusion the signals are often combined using statistical models such as Bayesian networks that incorporate probabilities of speech presence, obtained separately from each modality [94, 150]. In [47], we presented a kernel method, in which a low dimensional representation of the data is learned separately for each view. Then, two separate estimators for speech presence are constructed based on the two representations, and the data is fused by merging the estimators. In this chapter, we deviate from these two common approaches focusing on learning the complex mutual structures (relations) of the data in the two views.

In this study, we address the fusion problem of data obtained in multiple views. We revisit the alternating diffusion maps method and analyze it from

a different point of view than in [79] using a discrete setting. By adopting ideas from [42, 126], we study how connected data points on graphs of each view affect the connectivity of the graph obtained by fusing the views via the product of the affinity kernels. By assuming a statistical model on the connectivity between data points in each view, we use a simple argument to show that the kernel bandwidth of each view may be chosen such that the graph defined on each single view is not fully connected. This allows us to use significantly smaller kernel bandwidths improving the robustness of the fusion process to interferences. Based on the introduced statistical model, we propose an algorithm for the selection of the kernel bandwidth. We note that throughout this chapter we consider the selection of the kernel bandwidth for the alternating diffusion maps method presented in [79]. However, the provided analysis and the proposed algorithm may be extended with mild modifications to the methods presented in [42, 83, 90, 145], which are also based on the product between the affinity kernels of the views.

Using the alternating diffusion mas with the new algorithm for determining the kernel bandwidth, we address the problem of audio-visual voice activity detection, where the goal is to detect segments of the measured signal containing active speech. We consider a challenging setup in which a speech signal is measured by a single microphone and a video camera in the presence of high levels of acoustic noises and transients, which are short term interruptions, e.g., keyboard taps and office noise [46, 59]. In the video signal, there exist natural mouth movements during non-speech periods which wrongly appear similar to speech. The alternating diffusion maps method is particularly suitable for the fusion of the audio-visual data in this setup since it integrates out the interferences, which are view specific, i.e., transients measured in the microphone and non-speech mouth movements measured by the camera. Based on the alternating diffusion maps method, we propose a data-driven algorithm for voice activity detection. The algorithm comprises a simple preprocessing stage of feature extraction and does not require post-

processing, and, in contrast to the method we presented in [47], it requires no training data. Our simulation results demonstrate improved performance of the proposed algorithm compared both to a similar algorithm based on a traditional selection of the kernel bandwidth and compared to competing fusion schemes.

The remainder of the chapter is organized as follows. In Section 3.2, we briefly review the alternating diffusion maps method. In Section 3.3, we analyze the method in a discrete setting using tools from graph theory, and propose an algorithm for kernel bandwidth selection. In Section 3.4 we address the problem of audio-visual voice activity detection and propose an algorithm based on the alternating diffusion maps method. The improved performance of the proposed algorithm is demonstrated in Section 3.5.

3.2 Review of The Alternating Diffusion Maps Method

Consider a dataset of N samples captured in two different views given by:

$$(\mathbf{v}_1, \mathbf{w}_1), (\mathbf{v}_2, \mathbf{w}_2), ..., (\mathbf{v}_N, \mathbf{w}_N), \tag{3.1}$$

where $\mathbf{v}_n \in \mathbb{R}^{L_v}$ and $\mathbf{w}_n \in \mathbb{R}^{L_w}$ are the nth data points of the first and the second views, respectively. An example we will address under this setup is an audio-visual recording of a speaker, where $\mathbf{v}_n$ is the nth time frame of the signal captured in a microphone and $\mathbf{w}_n$ is the corresponding video frame of the mouth region of the speaker. The alternating diffusion maps method presented in [79] is a kernel based geometric method for data fusion. It is designed to reveal the geometric structure of the data, which is mutual to the two views ignoring the interferences, which are captured only in one of the views. In the following, we shortly describe the construction of alternating diffusion maps. Let $\mathbf{K}_v \in \mathbb{R}^{N \times N}$ be an affinity kernel representing affinities

between data points in the first view, such that the (n, m)th entry of the matrix, denoted by $K_v(n, m)$ is given by:

$$K_v(n, m) = \exp\left(-\frac{||\mathbf{v}_n - \mathbf{v}_m||^2}{\epsilon_v}\right), \tag{3.2}$$

where ϵ_v is the kernel bandwidth whose selection is discussed in details in Section 3.3. The affinity kernel $\mathbf{K}_v$ in (3.2) defines a graph on the dataset in the first view such that each data point is a vertex and $K_v(n, m)$ is the weight of the edge between vertex n and vertex m. Let $\mathbf{M}_v \in \mathbb{R}^{N\times N}$ be a row stochastic Markov matrix given by normalizing the rows of $\mathbf{K}_v$:

$$\mathbf{M}_v = \mathbf{D}_v^{-1}\mathbf{K}_v, \tag{3.3}$$

where $\mathbf{D}_v \in \mathbb{R}^{N\times N}$ is a diagonal matrix, whose nth element on the diagonal is denoted by $D_v(n, n)$ and is given by $D_v(n, n) = \sum_{m=1}^{N} K_v(n, m)$. In this study, we use a row normalization rather than a column normalization used in [79] allowing us to facilitate the discussion and results in Section 3.3, and our experimental results showed a negligible effect on the type of the normalization. The matrix $\mathbf{M}_v$ defines a Markov chain on the graph such that $\mathbf{M}_v(n, m)$ is the probability of transitioning from data point n to data point m in a single step. Similarly to $\mathbf{K}_v$, let $\mathbf{K}_w \in \mathbb{R}^{N\times N}$ be a matrix representing affinities between data points in the second view, and let $\mathbf{M}_w \in \mathbb{R}^{N\times N}$ be the corresponding row stochastic matrix. The views are fused by constructing a unified matrix, which is denoted by $\mathbf{M}$ and is given by the product of the row stochastic matrices [79]:

$$\mathbf{M} = \mathbf{M}_v \cdot \mathbf{M}_w. \tag{3.4}$$

The matrix $\mathbf{M}$ is also row stochastic and it integrates the relations between the data points over the two views; therefore, we term it the multiple view

Markov matrix. The continuous counterparts of the matrices $\mathbf{M}_v$ and $\mathbf{M}_w$ in (3.4) are typically considered in the literature as diffusion operators [34]. Likewise, the authors in [79] considered $\mathbf{M}$ as an alternating diffusion operator consisting of two diffusion steps on the two views, and showed that this alternating diffusion attenuates the view-specific interferences. In Section 3.4, we describe the construction of a unified low dimensional representation of the data through the eigenvalue decomposition of the matrix $\mathbf{M}$ similarly to obtaining a low dimensional representation of the data in a single view using principal component analysis.

3.3 Graph Theory Interpretation For Kernel Bandwidth Selection

Recall that the affinity kernel $\mathbf{K}_v$ in (3.2) defines a graph on $\{\mathbf{v}_n\}_{n=1}^N$ such that each data point $\mathbf{v}_n$ is a vertex and $K_v(m,n) = \exp\left(-\frac{||\mathbf{v}_n - \mathbf{v}_m||^2}{\epsilon_v}\right)$ is the weight of the edge between vertex n and vertex m. The kernel bandwidth ϵ_v controls the connectivity of the graph. When $||\mathbf{v}_n - \mathbf{v}_m||^2 < \epsilon_v$, high similarities are obtained between data points n and m, and they are considered connected; when $||\mathbf{v}_n - \mathbf{v}_m||^2 \gg \epsilon_v$ the similarity between the points is negligible and we assume no edge between the points. In order to capture the geometric structure of the data, common practice is to set the kernel bandwidth such that each data point is connected to at least one other point, i.e.:

$$\epsilon_v > \max_m \left[\min_n \left(||\mathbf{v}_n - \mathbf{v}_m||^2\right)\right]. \tag{3.5}$$

This choice is a necessary condition for the graph defined on the dataset to be connected such that there exists a path between every pair of points. In turn, a connected graph is a necessary condition for the eigenvectors of the affinity kernel to form a discrete orthogonal basis. This property is typically used for the construction of low dimensional representations [34, 143]. Yet,

the kernel bandwidth should be sufficiently small to prevent the association of data points with different content. In [67], the authors proposed choosing the value of the kernel bandwidth by:

$$\epsilon_v = C \cdot \max_m \left[\min_n \left(||\mathbf{v}_n - \mathbf{v}_m||^2 \right) \right], \tag{3.6}$$

where C is a parameter typically set in the range of $2 \div 3$ to empirically guarantee that the graph is connected in the single view case such that each point is connected to several other points. We note that the row normalization in (3.3) does not change the graph connectivity, but normalize the weights of each point such that they sum to one. In this chapter, we focus on the selection of the kernel bandwidth according to (3.6), yet other existing methods based, for example, on using a fixed number of connections to each point also rely on a similar graph connectivity [33, 143, 151].

In the multiple view case, the kernel bandwidth in each view is typically set in the literature as if the data is captured only in a single view and also require graph connectivity for each view, e.g, in [42, 79, 83, 145]. In contrast, we show that when the data is measured in multiple views, the graph of each single view does not necessarily have to be connected.

To demonstrate this idea, we consider a multiple view graph, which is defined by the Markov matrix $\mathbf{M}$ in (3.4). The vertices of the graph are pairs of data points $\{(\mathbf{v}_n, \mathbf{w}_n)\}_{n=1}^N$, and the matrix $\mathbf{M}$ defines a Markov chain on this graph such that the (n, m)th entry of $\mathbf{M}$ is the probability of a transition from vertex n to vertex m. For simplicity, we relate to (say) vertex n as to point n even though it is related to the pair of points $(\mathbf{v}_n, \mathbf{w}_n)$. The matrix $\mathbf{M}$ aggregates the relations between the data points based on the two views; there exists an edge between point n and point m in the multiple view graph if the transition probability between them, given by $\mathbf{M}(n, m)$, is non-zero.

To capture the geometric structure of the data, the necessary condition that each point is connected to at least one other point applies to the multiple view graph and not to the graphs of the single views. Namely, the single

view graphs can be disconnected while each point in the multiple view graph is connected[1] as demonstrated by Proposition 3.3. $\forall n$, $\exists m \neq n$ such that $M(n,m) \neq 0$ iff $\forall n$, $\exists m \neq n$ such that $M_v(n,m) \neq 0$ or $M_w(n,m) \neq 0$. Proposition 3.3 implies that each point in the multiple view graph is connected if it is connected at least in one of the views. If point n is disconnected in the first view, the nth row of the affinity kernel of the first view $\mathbf{K}_v$ is given by:

$$(K_v(n,1), K_v(n,2), ..., K_v(n,n), ..., K_v(n,N)) = (0,0,...,1,...0)\,.$$

Consequently, the nth row of the corresponding row stochastic Markov matrix $\mathbf{M}_v$ is given by:

$$(M_v(n,1), M_v(n,2), ..., M_v(n,n), ..., M_v(n,N)) = (0,0,...,1,...0)\,.$$

According to (3.4) and by the rule of matrix product, the nth row of $\mathbf{M}$ is given by:

$$(0,0,...,1,...0) \cdot \mathbf{M}_w = (M_w(n,1), M_w(n,2), ..., M_w(n,n), ..., M_w(n,N))\,.$$

Therefore, $M(n,m) \neq 0$ iff $M_w(n,m) \neq 0$. If point n is connected to (say) point m in the first view, i.e., $M_v(n,m) \neq 0$, by the matrix product rule, the nth row in $\mathbf{M}$ is given by a linear combination of row n and row m in $\mathbf{M}_w$; since $M_w(m,m) \neq 0$ (each point is connected to itself), we have that $M(n,m) \neq 0$. Namely, the necessary condition to learn the geometry of the data from two views is that each point is connected at least in one of the views. Therefore, the kernel bandwidths of each view may be set to small values without satisfying the requirement that the graphs of the single views are connected. We will show in Section 3.5 that assigning small values to the kernel bandwidth increases the robustness of the representation obtained

[1] We say that a point is connected if it is connected to at least one other point.

using the multiple view affinity kernel to interferences.

The remainder of this section revolves around the selection of the kernel bandwidth. By using a simplifying statistical model for the graph connectivity, we associate the selection of the kernel bandwidth in the single view case with the average number of connections, which we denote by δ. Then we show that in the multiple view case a proper kernel bandwidth is obtained by reducing the number of connections up to a root factor, i.e., $\sqrt{\delta}$. Based on this result, we present an algorithm for the selection of the kernel bandwidths.

Let $\mathbb{1}_v(n,m)$ be an indicator which equals one if point n and point m are connected in the first view and zero otherwise. For simplicity, we assume that each pair of data points is connected with probability p_v independently from all other data-points in the view. Namely, $\{\mathbb{1}_v(n,m)\}_{n,m}$ are independent and identically distributed (iid) random variables such that:

$$\mathbb{1}_v(n,m) = \left\{ \begin{array}{ll} 1, & \text{w.p. } p_v \\ 0, & \text{otherwise} \end{array} \right\}. \tag{3.7}$$

In addition, we assume that the connectivity between data points in a certain view is independent from the connectivity in the other views. We note that these two assumptions do not usually hold in practice. For example, two points being connected to a third point implies that the two points are close to each other, and as a result, they are connected with high probability. In addition, since the data from the different views are measurements of the same phenomenon, high correlation is expected across the views. Yet, we justify these assumptions by considering data contaminated with interferences, and assuming that the interferences reduce these correlations.

Based on this statistical model, the number of connections of a certain point to the other $N-1$ points in the graph of the first view is given by a

binomial distribution, denoted by B_v:

$$B_v\left(N-1, p_v\right).$$

The parameter p_v is directly related to the kernel bandwidth ϵ_v in (3.2); the larger the kernel bandwidth, the higher the probability that two points are connected. We assume that the kernel bandwidth, and therefore p_v, are chosen such that each point is connected on average to S_v points, i.e., $p_v \approx \frac{S_v}{N-1}$, because $p_v \cdot (N-1)$ is the expectation of the binomial distribution. Based on this model, the probability that a certain point is disconnected is denoted by q_v and is given by:

$$q_v = \left(1-p_v\right)^{N-1}.$$

For large values of N, we approximate q_v by:

$$q_v \approx \left(1 - \frac{S_v}{N-1}\right)^{N-1} \approx e^{-S_v}. \tag{3.8}$$

We note that we assumed in (3.8) that the average number of connections S_v does not depend on the number of data points N. In fact, some studies, e.g., the one presented in [151], suggest setting a constant number of connections to each point regardless to the size of the dataset.

In the single view case, the kernel bandwidth ϵ_v is chosen such that the graph is connected. Under this statistical model, it is equivalent to setting S_v such that the probability q_v in (3.8) approaches zero. Namely, setting the kernel bandwidth is equivalent to setting the average number of connections S_v to a certain value δ such that $e^{-\delta}$ approaches zero.

We proceed by considering the multiple view case, in which a similar statistical model is considered for the second view as well. Let p_w, S_w and q_w be the equivalents of p_v, S_v and q_v in the second view, respectively. We recall that a pair of points, point n and point m, is connected in the multiple view

graph if the (n,m)th entry of $\mathbf{M}$ in (3.4) is non-zero. The (n,m)th entry is explicitly written as:

$$M(n,m) = \sum_{l} M_v(n,l)\, M_w(l,m).$$

which implies that point n and point m are connected in the multiple view graph if there exists a third point l such that points n and l are connected in the first view and points l and m are connected in the second view. Accordingly, we show in the sequel that the pair (n,m) is connected with the approximated probability $\frac{S_v S_w}{N-1}$. The probability that the pair (n,m) is connected via a third point l, $l \neq n \neq m$, is $p_v p_w$, so there will be on average $(N-2)\,p_v p_w + p_v + p_w$ such connected triplets, where the two right terms correspond to the cases $l = m$ and $l = n$, respectively. We rewrite the term $(N-2)\,p_v p_w + p_v + p_w$ as:

$$(N-2)\,p_v p_w + p_v + p_w = (N-2)\,\tfrac{S_v}{N-1}\tfrac{S_w}{N-1} + \tfrac{S_v}{N-1} + \tfrac{S_w}{N-1} \approx \tfrac{S_v S_w}{N-1}$$

where the term $\frac{N-2}{N-1}$ approximately equals to one for large values of N, and the terms $\frac{S_v}{N-1}$ and $\frac{S_w}{N-1}$ have been neglected since $S_v S_w$ is larger than S_v and S_w by one order of magnitude. Typically, the average number of points connected to a certain point, S_v in the first view or S_w in the second view, is significantly smaller than N. Therefore, we assume that $\frac{S_v S_w}{N-1} < 1$, and we view this term as the probability that point n and point m are connected in the multiple view graph. The number of connections of each point in the multiple view graph is therefore given by the following binomial distribution:

$$B\left(N-1, \frac{S_v S_w}{N-1}\right), \tag{3.9}$$

and, specifically, each point in the multiple view graph is connected on average to $S_v S_w$ other points. Based on the binomial distribution and similarly to (3.8), the probability that a point is disconnected in the multiple view

graph, which we denote by q, is approximated by:

$$q \approx \left(1 - \frac{S_v S_w}{N-1}\right)^{N-1} \approx e^{-S_v S_W}.$$

We interpret this result similarly to the result obtained for the single view case; to meet the condition that each point in the graph is connected, the probability q has to approach zero, i.e., $q = e^{-\delta}$ as in the single view case, such that $S_v S_w = \delta$. Assuming for simplicity that $S_v = S_w$, the average number of connections in each view should be set to $S_v = S_w = \sqrt{\delta}$. In summary, in the multiple view case, we may reduce the average number of connections of each point by a root factor, and thus, significantly reduce the size of the kernel bandwidth, while meeting the requirement on the connectivity of the multiple view graph. We will show in Section 3.5 that this choice of the kernel bandwidth improves the representation obtained by the multiple view kernel method.

Next, we describe an algorithm for the selection of the kernel bandwidths. For simplicity, we consider the selection of the kernel bandwidth of the first view; the selection of the kernel bandwidth of the second view is equivalent. We start by estimating δ, i.e., the average number of connections to each point, $S_v = (N-1)\, p_v$, when the kernel bandwidth is selected according to (3.6) as if the data is captured only in a single view. Recalling that p_v is the probability that two arbitrary points are connected, we propose estimating it by:

$$\hat{p}_v = \frac{1}{N(N-1)} \sum_m \sum_{n \neq m} K_v(n,m), \tag{3.10}$$

where $\hat{p}_v$ is an estimate of p_v, and $K_v(n,m)$ is the (n,m)th entry of the affinity kernel $\mathbf{K}_v$ in (3.2). According to (3.2), $K_v(n,m)$ is in the range of $0 \div 1$ and a high value of $K_v(n,m)$ indicates that points n and m are connected. By selecting the kernel bandwidth according to (3.6) as if the data captured in a single view, the estimate of the average number of connections δ, denoted

by $\hat{\delta}$, is given by:

$$\hat{\delta} = (N-1)\,\hat{p}_v = \frac{1}{N}\sum_m \sum_{n\neq m} K_v\,(n,m)\,, \tag{3.11}$$

where we recall that $\delta = S_v = (N-1)\,p_v$. We denote the new bandwidth of the affinity kernel by ϵ_v^{AD}, where AD is alternating diffusion, and we select it such that the estimated average number of connections, which we denote by δ^{AD}, is reduced to $\sqrt{\hat{\delta}}$. We propose selecting ϵ_v^{AD} similarly to (3.6) by:

$$\epsilon_v^{\mathrm{AD}} = C^{\mathrm{AD}} \cdot \max_m \left[\min_{n\neq m}\left(||\mathbf{v}_n - \mathbf{v}_m||^2\right)\right], \tag{3.12}$$

where C^{AD} is a parameter in the range of $0 \div C$. The selection of a proper kernel bandwidth is reduced to the selection of the parameter C^{AD} decreasing the average number of connections by a root factor. We recall that to estimate $\hat{\delta}$ in (3.11), we choose $C = 2$ in (3.6) as if the data is captured in a single view, and propose searching the parameter C^{AD} within a discrete set whose elements lie on a linear grid. Let $\mathcal{C} = \{C_k\}_{k=1}^{|\mathcal{C}|}$ be a discrete set of size $|\mathcal{C}|$, where C_k, $k = 1, 2, ..., |\mathcal{C}|$ are given by $C_k = \frac{k}{|\mathcal{C}|}C$. We propose applying a binary search within the set such that the proposed kernel bandwidth, C^{AD}, is given by the element C_k for which the average number of connections δ^{AD} is the closest to $\sqrt{\hat{\delta}}$. We summarize the proposed algorithm in Algorithm 3.1.

3.4 Audio Visual Fusion With Application To Voice Activity Detection

We consider speech measured by a single microphone and by a video camera pointed to the face of the speaker. The audio-visual signal is processed in frames, and we consider a sequence of N frames, which are aligned in the two views (audio and video). While speech is measured in the two views,

Algorithm 3.1 Kernel bandwidth selection

1: Calculate ϵ_v in (3.6) by setting $C = 2$ as if the data is captured in a single view
2: Calculate $\mathbf{K}_v$ in (3.2)
3: Estimate $\hat{\delta}$ in (3.11)
4: Define: $\mathcal{C} = \{C_k\}_{k=1}^{|\mathcal{C}|}$, where $|\mathcal{C}| = 40$ and $C_k = \frac{k}{|\mathcal{C}|}C = 0.05k$
5: **while** $|\mathcal{C}| \neq 1$ **do**
6: $C^{\mathrm{AD}} = C_{|\mathcal{C}|/2}$
7: Estimate δ^{AD} similarly to (3.11) by recalculating $\mathbf{K}_v$ with the parameter C^{AD}
8: **if** $\delta^{\mathrm{AD}} > \sqrt{\hat{\delta}}$ **then**
9: $\mathcal{C} = \{C_k\}_{k=1}^{|\mathcal{C}|/2}$
10: **else**
11: $\mathcal{C} = \{C_k\}_{k=|\mathcal{C}|/2+1}^{|\mathcal{C}|}$
12: **end if**
13: **end while**
14: Using the obtained C^{AD}, calculate the new kernel bandwidth ϵ_v^{AD} in (3.12)

the interruptions are view specific. The audio signal consists of, in addition to speech, background noise and transients, which are short-term interferences, e.g. keyboard taps and office noise; the video signal contains mouth movements during non-speech intervals, which are considered as interferences since they appear similar to speech. Our goal is obtaining a representation of speech, which is robust to noise and interferences. In order to accomplish this goal, we apply alternating diffusion maps, where the kernel bandwidth is determined according to Algorithm 3.1.

The alternating diffusion maps method is applied in a domain of features, which are designed to reduce the effect of the interferences [47]. The audio signal is regarded as the first view, and it is represented by features based on Mel-Frequency Cepstral Coefficients (MFCC), which are commonly used for speech representation [85]. Specifically, the nth data point in the first view, $\mathbf{v}_n \in \mathbb{R}^{L_v}$ in (3.1), is the feature vector of the nth frame, and is given by the concatenation of the MFCCs of frames $n-1, n$, and $n+1$. Namely, L_v is

the total number of the coefficients in three consecutive frames. The use of consecutive frames reduces the effect of transients since speech is assumed more consistent over time compared to transients, which are rapidly varying. The data obtained in the second view, i.e., the video signal, is represented by motion vectors [22] such that the production of speech is assumed associated to high levels of mouth movement. The features representation of the nth frame of the video signal, $\mathbf{w}_n \in \mathbb{R}^{L_w}$, is given by concatenating the absolute values of the motion vectors in frames $n-1, n$ and $n+1$. Similarly to the audio signal, the use of consecutive frames for representation reduces the effect of short-term mouth movements during non-speech intervals. For more details on the construction of the features, we refer the reader to [47].

The representation using the specifically designed features is only partly robust to the interferences. For example, video features of a non-speech frame may be wrongly similar to the features of a speech frame if the former contains large movements of the mouth. To further improve the robustness of the representation to noise, the two views are fused using the alternating diffusion maps method with the improved affinity kernels. Specifically, we construct the affinity kernels of the two views, $\mathbf{M}_v$ and $\mathbf{M}_w$, according to (3.2) and (3.3) using the features $\{\mathbf{v}_n\}_{n=1}^N$and $\{\mathbf{w}_n\}_{n=1}^N$, respectively, and fuse the views by $\mathbf{M} = \mathbf{M}_v \cdot \mathbf{M}_w$. Then, we construct an eigenvalue decomposition of $\mathbf{M}$ such that the eigenvectors aggregate the connections between the data points within each view and between the views into a global representation of the data. Since the matrix $\mathbf{M}$ is row stochastic, the eigenvalue with the largest absolute value is 1 and it corresponds to an all ones eigenvector [34]. This eigenvector is neglected since it does not contain information. We note that the eigenvectors of $\mathbf{M}$ are not guaranteed to be real valued as it is guaranteed for the single view matrices $\mathbf{M}_v$ and $\mathbf{M}_w$ since the latter are similar to symmetric matrices. Therefore, one solution is using the singular value decomposition of $\mathbf{M}$; indeed, Lindenbaum et al. showed in [83] how to construct a new representation of the data using the singular value

decomposition of $\mathbf{M}$, in which the Euclidean distance between data points approximates a multiple view variant of the meaningful diffusion distance [34]. Yet, our experiments have shown that the eigenvectors corresponding to the several largest eigenvalues of $\mathbf{M}$ are indeed real and that the two approaches perform similarly.

We demonstrate the use of the representation obtained by alternating diffusion maps for the problem of voice activity detection. Let $\mathcal{H}_0, \mathcal{H}_1$ be hypotheses of speech absence and speech presence, respectively, and let $\mathbb{1}_n$ denote a speech indicator at the nth frame, given by:

$$\mathbb{1}_n = \left\{ \begin{array}{ccc} 1 & ; & \mathcal{H}_1 \\ 0 & ; & \mathcal{H}_0 \end{array} \right\}.$$

Given a sequence of N frames, the goal is to estimate the speech indicator, i.e., to separate the sequence of frames to speech and non-speech clusters. We found in our experiments that the obtained representation of the audio-visual signal, and specifically, its first coordinate, i.e., the leading (non-trivial) eigenvector of the matrix $\mathbf{M}$ in (3.4), which we denote by $\boldsymbol{\nu}_1 \in \mathbb{R}^N$, successfully separates between speech and non-speech frames. Therefore, we take a similar approach to [52] and estimate the speech indicator by comparing the leading eigenvector to a threshold τ:

$$\hat{\mathbb{1}}_n = \left\{ \begin{array}{ccc} 1 & ; & \nu_1(n) > \tau \\ 0 & ; & \text{otherwise} \end{array} \right\}, \tag{3.13}$$

where $\nu_1(n)$ in the nth entry of the eigenvector $\boldsymbol{\nu}_1$. We note that the leading eigenvector $\boldsymbol{\nu}_1$ is widely used in the literature for clustering and it was shown in [113] that it solves the well-known normalized cut problem. In contrast to previous works, in this study the leading eigenvector is obtained from the multiple view Markov matrix such that it clusters the data according to the two views. Similarly to [52], the leading eigenvector is used as a continuous measure of voice activity rather than for binary clustering. Therefore,

the threshold value controls the trade-off between correct detection and false alarm rates, and it may be chosen according to the specific application at hand. The proposed voice activity detection algorithm is summarized in Algorithm 3.2. Before proceeding to the experimental results, we note that the proposed representation and hence the speech indicator are obtained in a batch manner assuming that N consecutive frames are available in advance. Yet, as described in [47], a training set may be used to construct the representation, and then it can be extended to new incoming frames, e.g., using the Nyström method, in an online manner [55].

Algorithm 3.2 Voice activity detection

Calculate the features of the audio-visual signal $(\mathbf{v}_1, \mathbf{w}_1), (\mathbf{v}_2, \mathbf{w}_2), ..., (\mathbf{v}_N, \mathbf{w}_N)$
Calculate $\mathbf{K}_v$ and $\mathbf{K}_w$ according to (3.2) and Algorithm 3.1
Calculate $\mathbf{M}_v$ and $\mathbf{M}_w$ according to (3.3)
Fuse the views by calculating $\mathbf{M}$ in (3.4)
Obtain the leading eigenvector $\boldsymbol{\nu}_1$
for $n = 1 : N$ **do**
    if $\nu_1(n) > \tau$ **then**
        $\hat{\mathbb{1}}_n = 1$
    else
        $\hat{\mathbb{1}}_n = 0$
    end if
end for

3.5 Simulation Results

We use a dataset that we recently presented in [47]. The signals are recorded using a microphone and a frontal video camera of a smartphone pointed to the face of the speaker. The video signal is processed in 25 fps and it comprises the region of the mouth of the speaker automatically cropped out from the recorded video as described in [47] and illustrated in Fig. 3.1. The

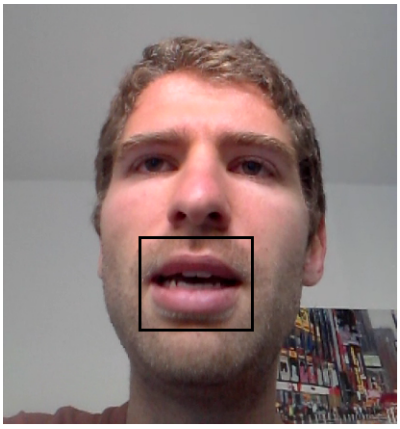


Figure 3.1: An example of a video frame and the cropped mouth region.

audio signal is processed in 8 kHz using time frames of 634 samples with 50% overlap, such that this setup aligns between the audio and the video signals. According to this type of alignment, the pair of points $(\mathbf{v}_n, \mathbf{w}_n)$ corresponds to the same time frame n, as required by the alternating diffusion maps method. In this context, we note that due to the different sampling rates, the raw audio data comprise samples of the measured phenomenon in different (finer) time scales than the video data. In this chapter, we neglect the miss-alignment in the measurement time that is below a single time frame, and consider the pair $(\mathbf{v}_n, \mathbf{w}_n)$ as two measurements obtained simultaneously. The dataset comprises 11 sequences of different speakers, each of which is 60 s long containing speech and non-speech intervals. Each of the 11 sequences is processed separately such that the number of frames in each sequence is $N \approx 1500$.

The signals are recorded in a quiet room and we synthetically add different types of background noise and transients to the audio signal. The transients are taken from an online free corpus [3] and they are normalized such that they have the same maximal amplitude as the clean audio signal. Based

on the clean audio, we mark the ground truth in each frame, such that frames with energy level higher than 1% of the highest energy level in the sequence are marked as speech frames. This setup of voice detection has a fine resolution of few tens of milliseconds, and it is useful for application such as speech recognition where single phonemes should be isolated [105, 106].

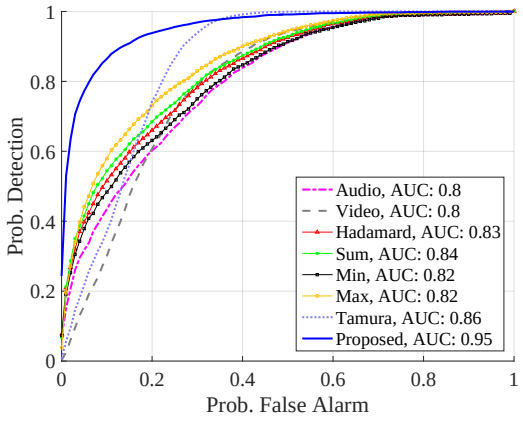
In the first experiment, we evaluate the proposed voice activity detection algorithm by comparing it to other versions of the algorithm based only on a single view (audio or video). We term the versions of the algorithm based on the first and the second view "Audio" and "Video", respectively, and the corresponding speech indicators are estimated by comparing the leading eigenvectors of the matrices $\mathbf{M}_v$ in (3.3) and $\mathbf{M}_w$ to a threshold, respectively. In addition, we examine another four approaches for the fusion of the views using the corresponding row stochastic matrices. In the first approach, the fused matrix is given by the Hadamard product between the matrices: $\mathbf{M}_v \circ \mathbf{M}_w$, where $\circ$ denotes point-wise multiplication; in the second approach, the views are fused by a simple sum: $\mathbf{M}_v + \mathbf{M}_w$; in the third and the forth approaches we use point-wise minimum and maximum functions. These approaches are termed in the plots "Hadamard", "Sum", "Min" and "Max", respectively. We note that both in the proposed algorithm and in the competing methods, the speech indicator is estimated by the leading eigenvector obtained by the eigenvalue decomposition with an arbitrary sign. To set the sign of the eigenvector, one may for example consider the variability of the video signal over time such that the lack of mouth movement over several consecutive frames indicates absence of speech. In this study, the sign of the eigenvector is assumed to be known for all the methods.

In addition to the different merging schemes, we compare the proposed algorithm to the method presented in [136] termed "Tamura" in the plots. The performances of the algorithms are presented in Fig. 3.2 for different types of transients in the form of receiver operating characteristic (ROC) curves, i.e., plots of probability of detection versus probability of false alarms.

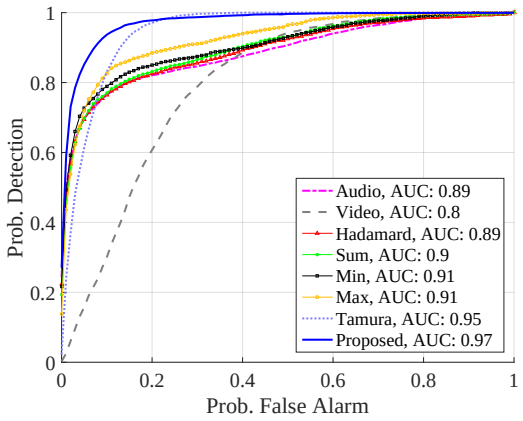
The larger the Area Under the Curve (AUC) is, the better the performance of the algorithms are, and the AUC of each algorithm is presented in the legend box. It can be seen in the plots that the algorithm based on the video signal provides relatively poor performance compared to the other algorithms. This is mainly since the ground truth is set to a fine resolution, and the video signal is not sensitive enough. For example, video frames of a closed mouth may be measured during both speech and non-speech intervals. We note that in most of the previous studies, the video signal is used for the detection of long speech intervals of several words, and it cannot detect speech in fine resolutions. In addition, we note that we also compared the proposed algorithm to the algorithm we recently presented in [47]. However, due to the challenging problem setting considered in this study, for which the speech is detected at a fine resolution, we found that incorporating the visual information as proposed in [47] does not improve the detection scores. Hence the simulation of [47] is not presented in the plots.

The audio signal in Fig. 3.2 also performs poorly due to the presence of transients, which are not properly separated from speech. The alternative fusion approaches, slightly benefit from the fusion of the sensors and provide performances comparable to the performance obtained by the audio signal. The proposed fusion of the audio-visual signal provides improved performance and outperforms all the other algorithms.

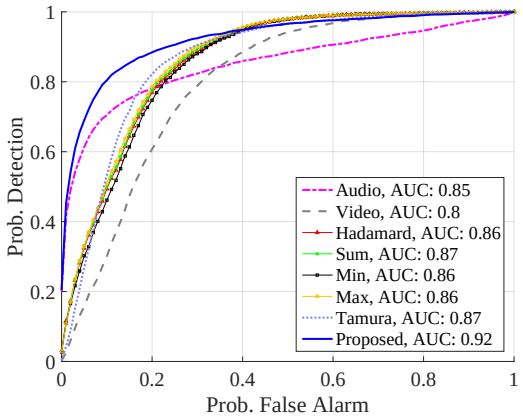
To further gain insight on the performance of the proposed algorithm for voice activity detection, we present in Fig. 3.3 an example of speech detection in a sequence contaminated by hammering. In this experiment, we set the threshold value in (3.13) to provide 90 percent correct detection rate and compare the false alarms resulting from the proposed algorithm to the false alarms resulting from the algorithm presented in [136]. As demonstrated in Fig. 3.3 (top), significantly less false alarms are received by the proposed



(a)



(b)



(c)

Figure 3.2: Probability of the detection vs probability of false alarm. Transient type: (a) hammering, (b) door-knocks, (c) microwave.

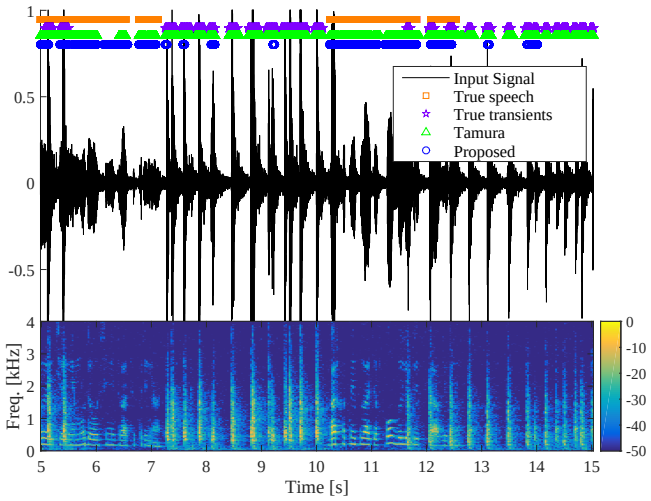


Figure 3.3: Qualitative assessment of the proposed algorithm for voice activity detection, with a hammering transient. (Top) Time domain, input signal- black solid line, true speech- orange squares, true transients- purple stars, "Tamura" with a threshold set for 90 percents correct detection rate- green triangles, proposed algorithm with a threshold set for 90 percents correct detection rate- blue circles. (Bottom) Spectrogram of the input signal.

algorithm compared to the competing detector such that the latter wrongly detects most of the transients as speech.

In Figs. 3.2 and 3.3, we calculate the parameters ϵ_v and ϵ_w for both the proposed and the alternative fusion approaches according to (3.6) by setting $C = 2$ as if the data is obtained in a single view. We note that also for the alternative fusion approaches, the necessary condition for the connectivity holds for the unified graphs defined by the corresponding fusion rules and not for the single view graphs. Accordingly, to properly select the kernel bandwidth for these approaches, further analysis of the connectivity of their corresponding graphs is required. Specifically, when $\epsilon_v = \epsilon_w$, the Hadamard approach is equivalent to concatenating each pair of data points $(\mathbf{v}_n, \mathbf{w}_n)$

into a column vector, which is regarded as a data point obtained in a single view (see Lemma 1 in [83]). In this case, the kernel bandwidth indeed may be selected as in the single view case; however, by setting $\epsilon_v = \epsilon_w$, the same weights are assigned to the distances between points in each view, which is not necessarily optimal since the data may be of different value range in the two views. Since in this study we focus only on the analysis of the kernel product, we find it convenient to compare the alternative fusion schemes by similarly selecting the kernel bandwidths of all the methods in a traditional manner as if the data is obtained in a single view.

We also evaluate the performance of the proposed algorithm for different values of the kernel bandwidth ϵ_v, and present the results in Fig. 3.4, where plots of the AUC of the proposed voice activity detector versus the parameter C in (3.6) for different types of noise and interferences are depicted. We recall that the parameter C represents the kernel bandwidth such that in the single view case, connected graphs typically correspond to C values in the range $2 \div 3$, and disconnected graphs correspond to C values less than 1. The red solid line in Fig. 3.4 is obtained by changing only the parameter C related to the audio signal while keeping the parameter related to the video signal fixed (with a constant value $C = 2$). The blue dot in the plots is C^{AD}, i.e., the proposed kernel bandwidth obtained by Algorithm 3.1. We empirically found that it is sufficient to search C^{AD} over a grid with a step size of $\frac{C}{|C|} = 0.05$, since tuning parameter values with larger accuracy showed negligible effect on the estimated average number of connections in the graph. It can be seen in the plots that by reducing the value of the parameter C the AUC is improved up to a peak obtained when $C \approx 0.5$. The peak value in the plots is the sweet spot in the trade-off in the kernel bandwidth selection. On the one hand, small values of the kernel bandwidth remove wrong connections in the graph between speech and non-speech frames, resulting in a representation in which these frames are better separated. On the other hand, too small kernel bandwidth causes the multiple view graph to be disconnected. Indeed,

the significant degradation of the AUC for parameter values below the peak may indicate that the multiple view graph is disconnected such that the obtained audio-visual representation no longer captures the geometric structure of speech. The fact that the peak is obtained for parameter value below 1 indicates that a better representation of the audio-visual signal is obtained by setting the kernel bandwidth such that the graph of the audio signal is disconnected. These plots demonstrate the idea that the kernel bandwidth should be chosen as the smallest possible keeping the graph of the multiple views connected. In addition, Fig. 3.4 demonstrate the performance of Algorithm 3.1 for the selection of the kernel bandwidth. The parameter C obtained by the algorithm, i.e., C^{AD}, successfully provides AUC close to the peak value.

The slight deviation of C^{AD} to the left of the peak may be explained by the assumptions on the statistical model in Section 3.3, which may not hold in practice. Specifically, the assumption that the connectivity in one view is independent of the connectivity in the other view does not hold in practice, since both views measure the same phenomenon. By taking the other extreme, assuming that the two views are fully dependent, i.e., the affinity matrices in the two views are identical, it may be shown in the continuous domain that the kernels product is equivalent to two diffusion steps of the size of the kernel bandwidth [75]. Namely, it is equivalent to multiplying the kernel bandwidth by a factor of two. Therefore, the kernel bandwidth in the multiple view case should be divided by two to maintain the connectivity as in the single view case. Accordingly, when the correlation between the views is not negligible, the proper kernel bandwidth should be set using the value of C in the range of $\left[C^{\mathrm{AD}}, 1\right]$, where $C = 1$ corresponds to the case of full correlation between the views. Since the maximum in Fig. 3.4 is obtained for a value of C, which is significantly smaller than 1, it implicitly implies a low correlation between the connectivity in the two views.

We note that we also applied Algorithm 3.1 for the selection of the kernel

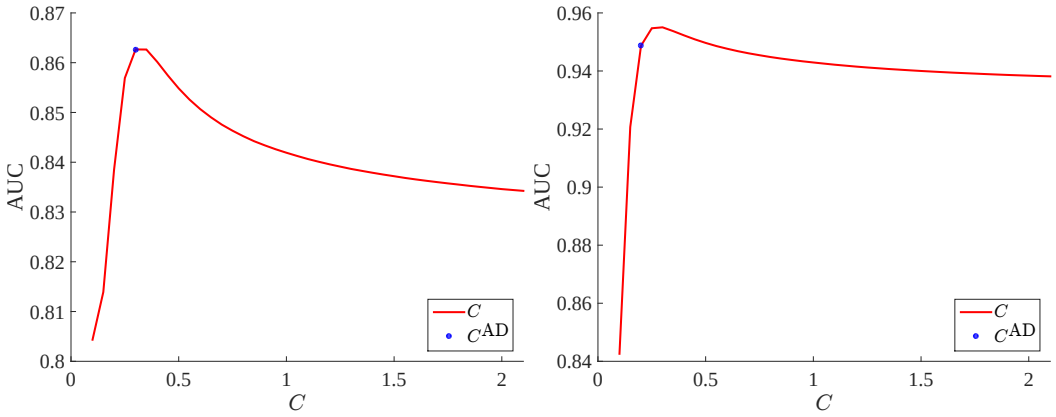


Figure 3.4: AUC vs the parameter C of the audio view. (a) background noise: babble noise with -5 dB signal to noise ratio (SNR), (b) transient type: keyboard-taps, background noise: colored Gaussian noise with -5 dB SNR.

bandwidth of the video signal. We found in our experiments that it performs comparably to the selection of the kernel bandwidth as in the single view case. Indeed, we expect to benefit from the algorithm only when there are high levels of noise and interferences. The video signal is considered relatively clean even though there exist some non-speech mouth movements, which may be wrongly detected as speech. In the case of clean signals, audio or video, there are significantly fewer wrong connections in the graphs of the single views, and hence, reducing the kernel bandwidths does not improve the obtained representation.

3.6 Conclusions

We have addressed the problem of multiple view data fusion. We revisited the alternating diffusion maps method and proposed a new interpretation from

a graph theory point of view, in which the affinity kernels of the single and multiple views define graphs on the data. By introducing a statistical model of the connectivity between data points on the graphs, we showed that fusing the data by a product of the affinity kernels increases the average number of connections in the multiple view case. Accordingly, the kernel bandwidth, controlling the connectivity between the data points, may be set significantly smaller than in the single view case. Specifically, we showed that the proper kernel bandwidth is the one reducing the average number of connections by a root factor, and presented an algorithm for its selection. Using the alternating diffusion maps method with the improved affinity kernel, we have addressed the problem of audio-visual fusion. In particular, we have considered the task of voice activity detection in the presence of transients; we have shown that the representation obtained by alternating diffusion maps allows for accurate speech detection using the first coordinate, i.e., the leading eigenvector. Our simulation results have demonstrated that the incorporation of visual data significantly improves the detection scores both compared to detecting speech based on only the audio data and compared to alternative merging schemes. In addition, our simulation results have demonstrated that reducing the kernel bandwidth below the values typically used in the single view case improves the robustness of the fusion to transient interferences and consequently the voice activity detection scores. Finally, we have demonstrated that the proposed algorithm for the kernel bandwidth selection allows for selecting near optimal values of the kernel bandwidth.

Chapter 4

Kernel Method for Speech Source Activity Detection in Multi-modal Signals

We consider a problem setup, in which a desired speech source is measured by a microphone and by a video camera in an interfering environment. We assume that the interfering sources in the audio signal are independent of the interfering sources in the video signal (e.g., the video signal does not capture the interfering speakers). Our objective in this chapter is to detect the activity of the desired source. To address this problem, we take a kernel based geometric approach for obtaining a representation of the measured signal, in which the effect of the interfering sources is reduced. Based on this representation, we devise a measure for the activity of the desired source; experimental results demonstrate its superiority compared to competing methods in the detection of speech signals in the presence of different challenging types of interferences, including interfering speakers in the audio signal.

4.1 Introduction

We address the problem of activity detection of a speech source, measured both by a microphone and by a video camera pointed at the face of the speaker. We term this source as "the desired source". Assuming that it is measured in the presence of interferences, the objective in this chapter is to detect the desired source while ignoring the interferences. We consider different types of interfering sources (interferences) in the audio signal, such as speech from other speakers, environmental noises, and transients, which are abrupt interruptions such as door-knocks [46, 47, 59]. The video signal may contain interferences such as head and mouth movements, which make the detection of the desired source difficult. Our main assumption is that the interferences in the two modalities (audio and video) are independent of each other, e.g., the video camera does not capture the interfering speakers.

The activity detection of the desired speech source may be useful for a variety of applications such as speech enhancement, speech and speaker recognition and speech diarization, where the goal is to determine "who spoke when" [29, 31, 32, 71, 95]. Speech diarization, for example, is a challenging problem since first, time intervals with active speech have to be accurately detected while ignoring both background noises, and transients, which often appear similar to speech [52], and second, the different speakers have to be distinguished, typically by assuming statistical models. In the audio-visual setting considered here, the activity of the desired source directly implies that the corresponding speaker is speaking regardless of presence or absence of interfering sources.

To address the problem of desired source activity detection, we take a multi-modal geometric approach, where the goal is to learn a representation of the data by exploiting relations (affinities) between data points in the different modalities (audio and video). Classical kernel based geometric methods, e.g., those presented in [13, 18, 34, 45, 110], typically address the problem of non-linear dimensionality reduction of single-modal data. They

are based on constructing an affinity kernel capturing relations between the data points, and provide a low dimensional representation via the eigenvalue decomposition of the affinity kernel. Recent studies suggest to extend these kernel based geometric methods to the multi-modal case by constructing separate affinity kernels for each modality, and then by fusing the modalities through different combinations of the affinity kernels, e.g., by their weighted sum [19, 20, 42, 61, 72, 73, 79, 81–83, 90, 145, 153].

Lederman and Talmon presented in [79] a multi-modal fusion approach, where the data in the different modalities is fused by a product of affinity kernels, constructed separately for each modality. This fusion approach is particularly useful for the representation of the desired audio-visual source since, according to the analysis presented in [79], it reduces the effect of modality-specific sources, which in our problem setting are the interferences, by assumption. Hence, the obtained representation respects the relations between the data points according to the source present in both the modalities, which is the desired source in our case; therefore, it is particularly suitable for the activity detection of the desired source. In [48], we analyzed this fusion approach in a discrete setting showing that it may be further improved by a proper selection of the kernel bandwidth.

In this chapter, we propose an algorithm for activity detection of a desired speech source. The algorithm is based on constructing two affinity kernels, one for each modality (audio and video), in a domain of features, separately built for each modality. We fuse the modalities by a product of the affinity kernels as in [48, 79] and devise a measure for the presence of the desired source using the eigenvalue decomposition of the product kernel. We apply the proposed algorithm for the detection of audio-visual speech signals in the presence of multiple interfering audio sources including different speakers, background noises, and transients. Our simulation results demonstrate improved detection scores compared to single-modal variants, which are based on either the audio or the video signals, as well as compared to alternative

fusion schemes.

We note that we consider as the main challenge in this study, the presence of *multiple* interfering sources. Specifically, we consider interferences that are of the same type as the desired source, i.e., other speakers in the audio signal. In addition, the video signal comprises the entire face of the speaker; therefore, head movements are considered interferences in the video. We note that in [48], we addressed a special case of the problem that is considered here; previously, we considered the presence of only a single interfering transient noise source, which is considerably different from speech. In addition and in contrast to this chapter, only the mouth region of the speaker was assumed as the video signal, requiring an accurate detection of the mouth region as a preprocessing stage.

The remainder of the chapter is organized as follows. In Section 4.2, we formulate the problem and in Section 4.3 we present the proposed algorithm for speech source activity detection. The improved performance of the proposed algorithm is demonstrated in Section 4.4.

4.2 Problem Formulation

Consider a speech signal measured by a single microphone and by a video camera pointed at the face of a speaker. The signal is processed in consecutive frames, which are assumed aligned; let $\mathbf{v}_n \in \mathbb{R}^{L_v}$ and $\mathbf{w}_n \in \mathbb{R}^{L_w}$ be feature representations of the nth time frame in the first and the second modalities (i.e., audio and video), respectively, such that L_v and L_w are the total number of features in each modality. We use the MFCC [85] and motion vectors [22], for the representation of the audio and the video signals, respectively, as we describe in detail in [47]. The MFCCs are widely used for the representation of audio signals, and the motion vectors capture the movement of the mouth within the video, assumed to be associated with speech. In both modalities, we aggregate the features of three consecutive frames such that $\mathbf{v}_n$ is given

by the MFCCs of frames $n-1, n, n+1$. Consider a sequence of N such pairs of frames:

$$\{(\mathbf{v}_n, \mathbf{w}_n)\}_{n=1}^{N}. \tag{4.1}$$

We assume that the measured audio signal comprises $M^v + 1$ sources: $S_1, S_2, ...S_{M^v}$ and S^d, where the superscript d stands for the desired source. Namely, the audio frame $\mathbf{v}_n$ is given by a mapping, denoted by f, of the sources to the features space:

$$\mathbf{v}_n = f(S^d, S_1, S_2, ...S_{M^v}).$$

The video signal comprises the video recording of the face of a speaker. Yet, there may be both natural mouth and head movements, which are not directly related to speech and are considered as interferences. Assuming M^w such interfering sources, the corresponding video frame $\mathbf{w}_n$ is given by:

$$\mathbf{w}_n = g(S^d, S_1, S_2, ...S_{M^w}),$$

where g denotes the mapping of the sources to the feature space of the video signal. With the exception of the desired source, the sources of the audio and the video signals are assumed independent. In addition, the sources are assumed to be present or absent independently of each other. Specifically, we assume two hypotheses, $\mathcal{H}_0$ and $\mathcal{H}_1$, for the absence and the presence of the desired source, respectively. Accordingly, let $\mathbb{1}_n$ be an indicator for the presence of the desired source in the nth frame, given by:

$$\mathbb{1}_n = \left\{ \begin{array}{lcl} 1 & ; & n \in \mathcal{H}_1 \\ 0 & ; & n \in \mathcal{H}_0 \end{array} \right\}. \tag{4.2}$$

The goal in this study is to detect the activity of the desired source, i.e., to estimate the indicator in (4.2).

4.3 Desired Speech Source Activity Detection

4.3.1 Multi-modal Fusion via the Product of Affinity Kernels

For completeness, we describe the fusion process based on a product between affinity kernels constructed separately for each modality, as proposed in [79]. Let $\mathbf{K}_v \in \mathbb{R}^{N\times N}$ be an affinity kernel of the first modality (i.e., audio), whose (n,m)th entry, denoted by $K_v(n,m)$, is given by:

$$K_v(n,m) = \exp\left(-\frac{||\mathbf{v}_n - \mathbf{v}_m||^2}{\epsilon_v}\right), \tag{4.3}$$

where ϵ_v is the kernel bandwidth whose selection we studied in [48]. By dividing each column by its sum, we construct a row stochastic matrix, which is denoted by $\mathbf{M}_v \in \mathbb{R}^{N\times N}$, and its (n,m)th entry, $M_v(m,n)$, is given by:

$$M_v(n,m) = \frac{K_v(n,m)}{d_v(n)}, \tag{4.4}$$

where $d_v(n) = \sum_{m=1}^{N} K_v(n,m)$. Similarly to $\mathbf{M}_v$, we construct a row stochastic matrix $\mathbf{M}_w \in \mathbb{R}^{N\times N}$ for the second modality, and the data from the two modalities are fused by the product of the row stochastic matrices:

$$\mathbf{M} = \mathbf{M}_v \cdot \mathbf{M}_w, \tag{4.5}$$

where $\mathbf{M} \in \mathbb{R}^{N\times N}$ is viewed as aggregating the relations between the data points in the two modalities. Lederman and Talmon considered in [79] the continuous counterparts of $M_v(n,m)$, $M_w(n,m)$ and $M(n,m)$ as diffusion operators. They showed that the continuous operator corresponding to $M(n,m)$ is an alternating diffusion operator, which integrates out modality-specific sources by applying the diffusion process in two steps corresponding to the two modalities.

4.3.2 Desired Source Activity Detection

For the detection of the desired source, we apply an eigenvalue decomposition to $\mathbf{M}$. The eigenvectors respect the relations between the multi-modal data points aggregated in the matrix $\mathbf{M}$, and therefore, they are often used in the literature to form a low dimensional representation of the data [34]. The matrix $\mathbf{M}$ is row stochastic since $\mathbf{M}_v$ and $\mathbf{M}_w$ are row stochastic matrices, so it has an all ones eigenvector corresponding to the eigenvalue one, which we neglect since it does not contain information [34]. Since $\mathbf{M}$ integrates out the modality-specific sources, which are the interferences in our case, its eigenvectors represent the data according to the desired audio-visual source. For the detection of the desired source, we use the leading (non-trivial) eigenvector, which we denote by $\boldsymbol{\nu}_1 \in \mathbb{R}^N$; the nth entry of $\boldsymbol{\nu}_1$, denoted by $\nu_1(n)$, corresponds to the nth frame of the measured signals (audio and video) and we view it as a new mapping h of the nth frame according to the desired source:

$$\nu_1(n) = h(S^d).$$

The leading eigenvector of an affinity kernel is typically used in the literature for clustering such that the nth data point is clustered according to the sign of $\nu_1(n)$ [113]. Indeed, we have found in our experiments, that the data are properly clustered by $\boldsymbol{\nu}_1$ according to the presence and the absence of the desired source. Accordingly, we propose to estimate the indicator for the presence of the desired source $\mathbb{1}_n$ in (4.2) by comparing the eigenvector entries to a threshold τ:

$$\hat{\mathbb{1}}_n = \left\{ \begin{array}{lcl} 1 & ; & \nu_1(n) > \tau \\ 0 & ; & \text{otherwise} \end{array} \right\}. \tag{4.6}$$

Namely, we view the leading eigenvector as a continuous measure of the presence of the desired source. The threshold τ controls the trade-off between the probability of correct detection of the desired source and the probability

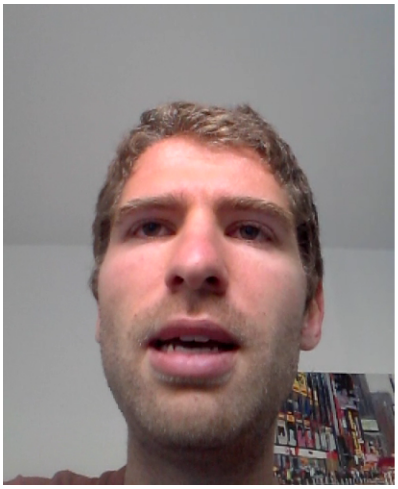


Figure 4.1: An example of a video frame.

of false alarm, and its setting is application dependent.

4.4 Experimental Results

We consider an audio-visual recording of a speaker measured by a microphone and by a video camera pointed at the face of the speaker. We use a dataset, which we presented in [47], comprising 11 sequences of different speakers, 60 s long each. The video signal is measured in 25 fps frame rate, and the audio signal, which is measured in 8 kHz, is aligned to the video signal using frames of 634 samples with 50% overlap. To simulate the interferences, we synthetically add to the audio signal different types of background noises and transients taken from a free online corpus [3], and other speakers taken from the dataset in [47]. The video signal comprises the entire face of the speaker as demonstrated in Fig. 4.1, in contrast to [48], where cropping of the mouth region of the speaker was required as a preprocessing step. Therefore, it may contain natural head and mouth movements, which are not related to

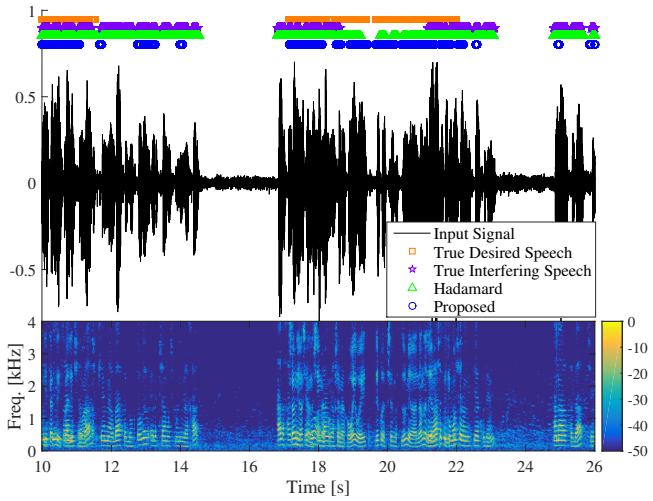


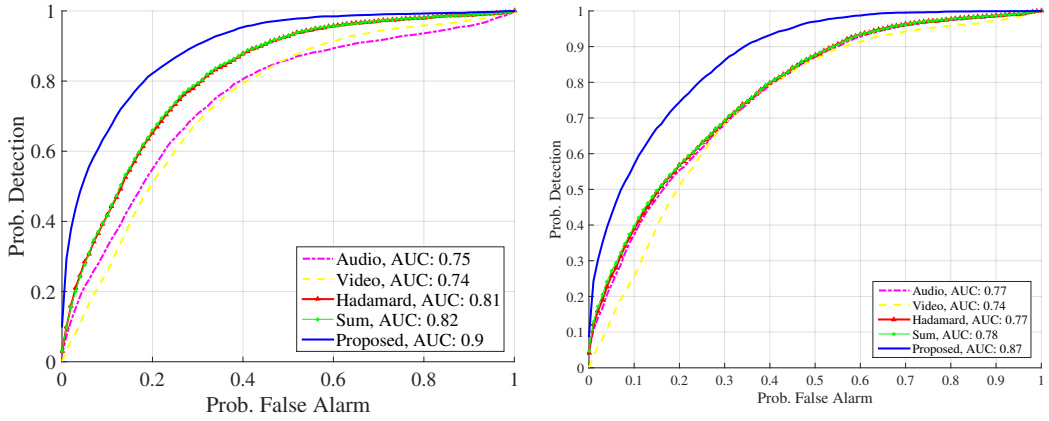
Figure 4.2: Qualitative assessment of the proposed algorithm for the desired speech source activity detection in the presence of three sources: the desired speech source, an interfering speech source, and babble noise with 20 dB SNR. (Top) Time domain, input signal - black solid line, true desired speech source - orange squares, true interfering speech source - purple stars, "Hadamard" with a threshold set for 80% correct detection rate - green triangles, proposed algorithm with a threshold set for 80% correct detection rate- blue circles. (Bottom) Spectrogram of the input signal.

speech. To set the ground truth of the activity of the desired speech source (which also appears in the video), we use the clean audio signal and consider the desired source active in a frame if its energy level is above 1% from the maximal energy value in the sequence. In this type of ground truth setting, the resolution of the presence and absence of the desired source is up to a single frame and it may be used, for example, for the enhancement of the desired source [29].

An example of the detection of the desired speech source obtained by the proposed algorithm is presented in Fig. 4.2, where we consider three audio sources comprising two speakers – one desired, one interfering and babble

noise. For the clarity of presentation, we use a relatively high SNR of 20 dB, where the SNR is calculated with respect to the desired speaker and the babble noise. Hence, the main challenge in this example is to distinguish between the desired speech signal and the speech signal of the interfering speaker. Indeed, the spectrogram of the measured audio signal, presented in Fig. 4.2 (Bottom), demonstrates that just by observing the spectrogram, it is hard to distinguish between the speech parts corresponding to the desired speech and the interfering speech. In Fig. 4.2 (Top), we qualitatively compare the proposed method for the detection of the desired source to an alternative kernel method termed "Hadamard", in which, instead of the product between the kernels in (4.5), the modalities are fused by the Hadamard product: $\mathbf{M}_v \circ \mathbf{M}_w$, where $\circ$ denotes point-wise multiplication. For both approaches, we set the value of the threshold τ in (4.6) to provide 80% correct detection rate and compare their false alarm rates. It may be seen that the proposed approach provides significantly fewer false alarms, and the competing method wrongly detects the activity of the interfering speech source, e.g., in the time interval after the 24th second.

In Fig. 4.3 we present the results of a quantitative evaluation of the proposed approach in the form of ROC curves, which are plots of detection versus false alarm rates. The proposed approach is compared, in addition to the method "Hadamard", to a method based on fusing the modalities via a sum of the affinity kernels, i.e., $\mathbf{M}_v + \mathbf{M}_w$, termed "Sum" in the plots. In addition, we compare the proposed approach to its single-modal variants, termed "Audio" and "Video", which are based on estimating the speech indicator in (4.6) using the leading eigenvector of the kernels $\mathbf{M}_v$ and $\mathbf{M}_w$, respectively. We observe in the plots that the approaches based on a single modality attain comparable results; the detector of the desired source based only on the audio signal is limited due to high similarity of the desired source especially to the other speakers. The performance based only on the video signal are also limited both due to modality-specific sources such as



(a) (b)

Figure 4.3: Probability of the detection vs probability of false alarm. Source types: (a) two speakers and babble noise with 20 dB SNR, (b) two speakers, door-knocks transients and white Gaussian noise with 15 dB SNR.

movements of the head and due to the high resolution of the ground truth. Indeed, there exist speech parts that do not involve the movement of mouth in certain time frames. The alternative fusion schemes perform slightly better than the single-modal approaches. Finally, the proposed approach for the detection of the desired source outperforms all other methods and provides improved detection scores for all false alarm values.

4.5 Conclusion

We have addressed the problem of audio-visual speech source detection in the presence of interferences. We proposed an algorithm for the detection of a desired source by fusing the modalities via a product of kernels, constructed separately for each modality. An eigenvalue decomposition of the product

kernel yields a useful representation of the data, in which the effects of the interfering sources are reduced, allowing us to devise a measure of the presence of the desired source based on the leading eigenvector. Experimental results have demonstrated the improved performance of the proposed algorithm in challenging environments, including speech activity detection in audio-visual data under presence of modality-specific interfering sources.

Chapter 5

Multi-modal Kernel Method for Activity Detection of Sound Sources

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 chapter is to detect the activity of the source of interest even when the video data is 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.

Acoustic scene, data fusion, multi-modal, audio-visual, transient noise, kernel

5.1 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 chapter, 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 [52]. 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. 5.1, 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

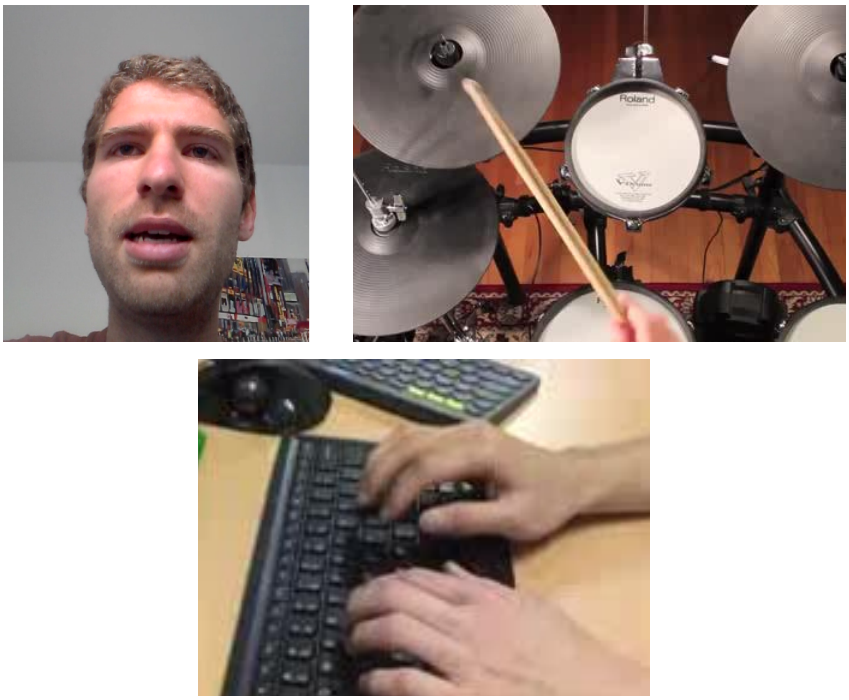


Figure 5.1: Examples of video frames of sources of interest. From left to right: speech, drum beats, keyboard-tapping.

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 [31, 133], which is carried out in this chapter 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 [28, 37, 43, 65, 124, 139, 140]. The present work belongs to a recent line of studies dealing with a polyphonic setting, where multiple sounds may be active simultaneously [4, 12, 23, 76, 88, 101]. There are several 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 [118] 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 [128], 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 [24, 54, 63, 69, 111, 112]. In [63],

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 [54], where the authors suggested to map audio and video signals into domains designed to maximize the mutual information between the modalities. The authors in [69] 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 [111, 112] 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 [24, 54, 63, 69, 111, 112] 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 [13, 18, 34, 45, 110]. 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 [19–21, 42, 48, 61, 72, 73, 79, 81, 83, 135, 145, 153].

A particular data fusion approach, which is based on combining the data via the product of affinity kernels, was recently studied in [48,79,135]. In [48], 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 [79], 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 chapter, 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 [48, 79], 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 chapter with respect to our previous work presented in [48] are 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 [48]. As far as we know, this chapter is the first to demonstrate a successful extension of the fusion method presented in [48, 79] to partially available multi-modal signals, i.e., signals measured by sensors of different types (audio and video). In addition, in [48], 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 chapter, 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 [84, 100, 117], 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 chapter is organized as follows. In Section 5.2, we formulate the problem. In Section 5.3, we propose an algorithm for activity detection of sources of interest, and present experimental results demonstrating its improved performance in Section 5.4.

5.2 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 5.4. 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 5.4. 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 5.3 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:

$$\mathbb{1}_n = \left\{ \begin{array}{ll} 1, & n \in \mathcal{H}_1 \\ 0, & n \in \mathcal{H}_0 \end{array} \right\}. \tag{5.1}$$

The goal in this chapter is to detect the presence of the source of interest, while ignoring all other sources, i.e., to estimate $\mathbb{1}_n$ in (5.1). 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 (5.1) such that even for the interval $1, 2, ..., L$ we do not have labels indicating the presence of the sources.

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

5.3 Kernel-based Detection of the Source of Interest

5.3.1 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:

$$K^a_{n,m} = \exp\left[-\left\|\mathbf{a}_n - \mathbf{a}_m\right\|_2^2 / \epsilon^a\right], \tag{5.2}$$

where $\|\cdot\|_2$ is the L_2 distance, and ϵ^a is the kernel bandwidth, a parameter whose selection we studied in [48]. The affinity kernel has an interpretation of a graph on the data, which we term the audio graph, whose nodes are the data points $\{\mathbf{a}_n\}$, 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:

$$D^a_{n,n} = \sum_{m=1}^{L} K^a_{n,m}. \tag{5.3}$$

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:

$$\mathbf{M}^a = \left(\mathbf{D}^a\right)^{-1}\mathbf{K}^a. \tag{5.4}$$

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 [34], 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:

$$\mathbf{M} = \mathbf{M}^a\mathbf{M}^v, \tag{5.5}$$

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 (5.2)-(5.3). 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 (5.5), 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 [79] 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 5.3.2, we provide more insight on this result by describing the relation between the product of kernels and the diffusion distance [34], which in turn motivates us to build a measure for the presence of the source of interest as we describe

in Subsection 5.3.3.

5.3.2 Diffusion Distance

Let $d(n,m)$ be the diffusion distance between frame n and frame m, given by [79]:

$$d(n,m) = \sqrt{\sum_{l=1}^{N} (M_{n,l} - M_{m,l})^2}. \tag{5.6}$$

According to (5.6), 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 [34], 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 [34]. 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, $s_1^a, s_2^a, ..., s_R^a, s_1^v, s_2^v, ..., s_Q^v, \tilde{s}$.

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:

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

where T denotes transpose, and $\mathbf{h}_n^0 \in \mathbb{R}^L$ is an indicator vector whose nth element equals one and all other elements equal zero. Accordingly, the diffusion distance $d(n,m)$ in (5.6) is given by:

$$d(n,m) = \|\mathbf{h}_n - \mathbf{h}_m\|_2 . \tag{5.7}$$

The use of the product of kernels for the fusion of the audio and the video

signals is motivated by Theorem 5 in [79], 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 (5.6). 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 [79].

5.3.3 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 (5.5). 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 [34, 79]:

$$d\left(n,m\right) = \sqrt{\sum_{l=1}^{N} \lambda_l \left(\phi_l(n) - \phi_l(m)\right)^2}, \tag{5.8}$$

where $\phi_l(n)$ is the nth entry of $\boldsymbol{\phi}_l$. The expression in (5.8) 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 (5.1), by:

$$\hat{\mathbb{1}}_n = \left\{ \begin{array}{lcl} 1 & ; & \phi_1(n) > \tau \\ 0 & ; & \text{otherwise} \end{array} \right\}, \tag{5.9}$$

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 [113]. 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 5.4, 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 [113].

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 (5.9), 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 [90, 135], as we describe next. Given

a new frame $\mathbf{a}_n$, $n \in [L+1, L+2, ..., N]$, we use the nyström method [55] to obtain a new entry of $\boldsymbol{\phi}_1$ corresponding to frame n, which is denoted by $\phi_1(n)$:

$$\phi_1(n) = \frac{1}{\lambda_1} \sum_{m=1}^{L} M_{n,m} \phi_1(m). \tag{5.10}$$

By (5.5), (5.10) can be rewritten as:

$$\phi_1(n) = \frac{1}{\lambda_1} \sum_{m=1}^{L} \sum_{l=1}^{L} M^a_{n,l} M^v_{l,m} \phi_1(m) \triangleq \sum_{l=1}^{L} M^a_{n,l} \theta(l), \tag{5.11}$$

where $\theta(l) = \frac{1}{\lambda_1} \sum_{m=1}^{L} M^v_{l,m} \phi_1(m)$. The right term in (5.11) 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 $1, 2, ..., L$.

Algorithm 5.1 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 (5.2)
Calculate the row stochastic matrices $\mathbf{M}^a$ and $\mathbf{M}^v$ according to (5.3)-(5.4)
Fuse the data via the product of kernels, i.e., compute $\mathbf{M}$ according to (5.5)
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 (5.11)
    if $\phi_1(n) > \tau$ **then**
        $\hat{\mathbb{1}}_n = 1$
    else
        $\hat{\mathbb{1}}_n = 0$
    end if
end for

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 [34], 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 [83] 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 5.1.

5.4 Experimental Results

5.4.1 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 are processed in an online manner and are 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 ~ 630 samples such that the audio frames are aligned with the video frames. To evaluate

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

the performance 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 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 [41, 58, 85]. We use a Matlab implementation of the MFCCs, taken from [2], 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" signal, 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 [39]. 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 [15, 22] (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 (5.11), a more computationally efficient extension is obtained by:

$$\phi_1(l) = \sum_{m=1}^{L} M^a_{l,m} \phi_1(m), \tag{5.12}$$

The extension in (5.12) 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 (5.12) 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 [99], which is based on using eigenvectors of an affinity kernel for speech detection.

In addition, we compare the proposed algorithm to the CCA method, which is denoted by "CCA" in the plots, and to the method presented in [54]. 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 [54] 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.

5.4.2 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 presented in [47] 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 [3], 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 [52]. 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. 5.2, 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. 5.2 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. 5.2 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 (5.9) to provide 80% correct detection rate and compare their false alarms.

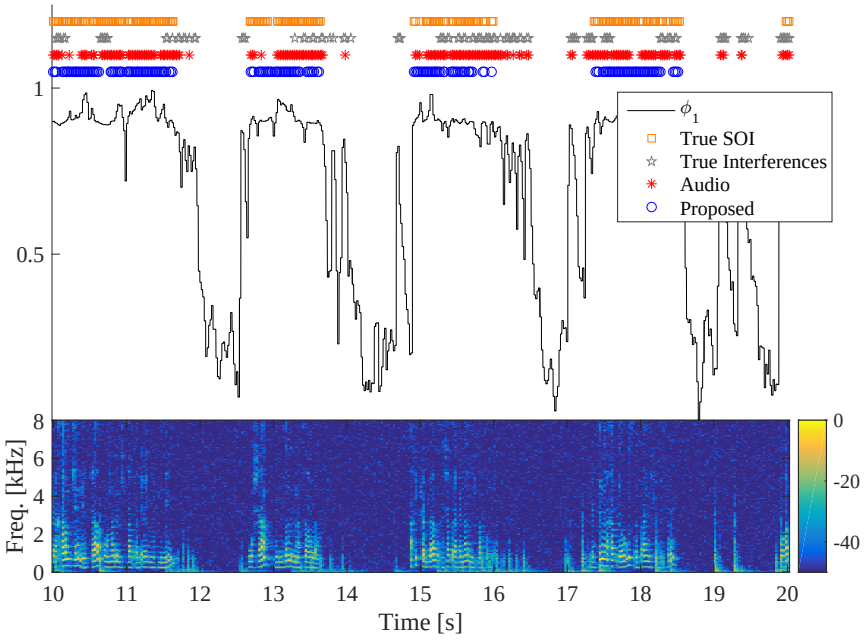


Figure 5.2: Qualitative assessment of the proposed algorithm for the activity detection of the source of interest. Source of interest: speech. Interfering source: door-knock transients with SIR 1. (Top) Time domain, trajectory of the leading eigenvector - black solid line, true SOI (speech) - orange 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.

It can be seen in Fig. 5.2 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. 5.3 in the form of 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 (5.9) 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 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. 5.3 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 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 (5.11) 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 (5.12) 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 [48] we considered the fusion of audio-visual data using the product of kernels for speech detection. We showed that it

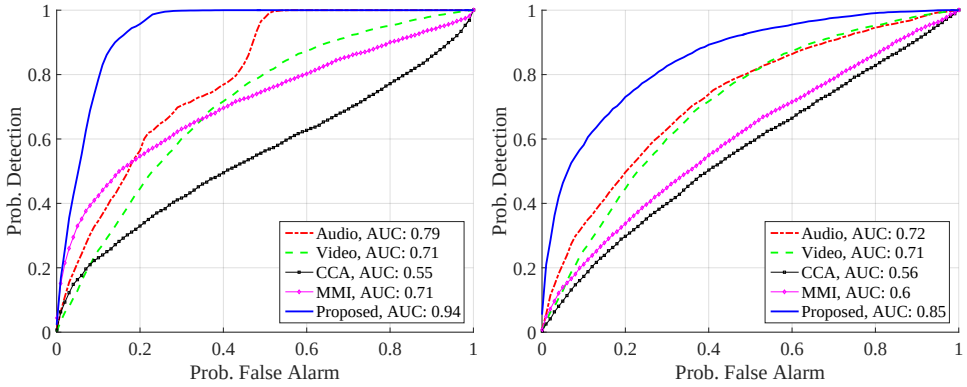


Figure 5.3: Probability of detection vs probability of false alarm. Source of interest: speech. Interfering sources: (a) door-knock transients with SIR 1, (b) babble noise with 0 dB SNR and scissors transient with SIR 1.

provides better detection scores compared to alternative fusion schemes and the methods presented in [47, 136]. However, in [48], 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 [48], 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. Figure 5.3 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 5.1 (a-c) for different SIR levels. Table 5.1 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 5.1 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.

5.4.3 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. 5.4 and 5.5, 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 [31] to reduce stationary noise. As an interfering source in these experiments, we use, in addition to transients, speech signals taken from TIMIT database [56].

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

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 interests and interference
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)

Table 5.1: (a-c) AUC scores. 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.

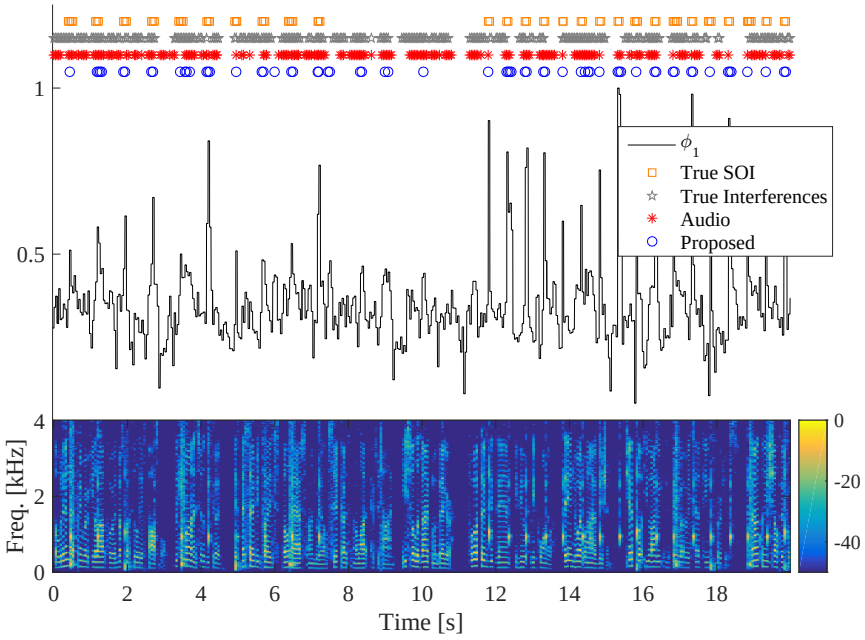


Figure 5.4: Qualitative assessment of the proposed algorithm for the activity detection of the source of interest. Source of interest: drum beats. Interfering source: speech with SIR 2. (Top) Time domain, trajectory of the leading eigenvector - black solid line, true SOI (drum beats) - orange squares, 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.

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. 5.1.

Figure 5.4 demonstrates the accurate detection of drum beats in the presence of interfering speech. We consider the drum beats as an example of chal-

lenging 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 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. 5.2. 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. 5.5, 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

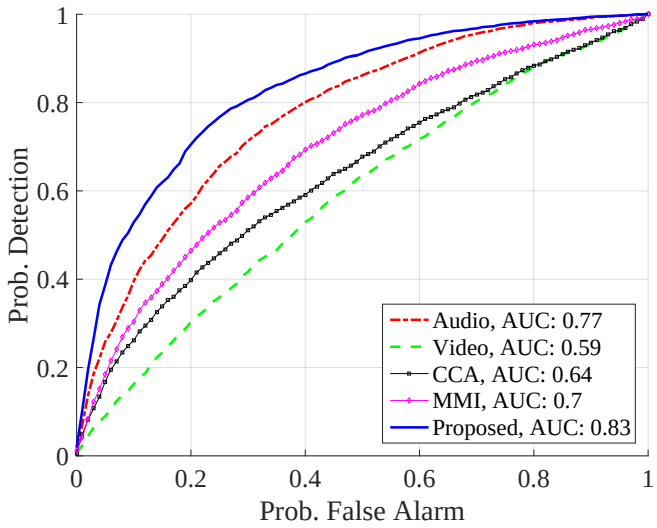


Figure 5.5: Probability of detection vs probability of false alarm. Source of interest: keyboard taps. Interfering source: speech with SIR 2.

the alternative approaches.

In Table 5.2 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 using the alternating diffusion maps method for improving the analysis of complex sound scenes.

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)

Table 5.2: (a-c) AUC scores. 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.

5.4.4 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 [5] 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 [74]. In the latter case, labels in the testing set are estimated by iteratively diffusing training labels with the testing set according to similarities (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.

5.5 Conclusions

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.

Chapter 6

Sequential Audio-visual Correspondence With Alternating Diffusion Kernels

A fundamental problem in multi-modal signal processing is to quantify relations between two different signals with respect to a certain phenomenon. In this chapter, 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.

6.1 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 chapter, 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 chapter.

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 [13, 18, 34, 45, 110]. 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 [19–21, 42, 48, 61, 72, 73, 79, 81, 83, 89, 145, 153]. 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 [48, 79, 89]. Lederman and Talmon analyzed the kernel product in a continuous setting, showing that it recovers the common components from multi-modal observations. In [48], 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 [89] the equivalence of this fusion approach to a non-parametric variant of kernel CCA.

Here we consider the correspondence between multi-modal signals, which was not previously addressed in [19–21, 42, 48, 61, 72, 73, 79, 81, 83, 145, 153]. 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 [89, 96], and then extended to other samples using, e.g., Nyström method [55]. In this context, we mention the studies presented in [44, 80, 147], 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 [35, 36, 91–93, 103, 125, 144]. 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 chapter. Izadinia et al. [64] 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. [92] 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. [152], 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 problem of audio-visual localization was also addressed in [16, 68, 69, 98], typically formulated as an optimization problem for learning unified representations of the audio-visual signals. For example, Kidron et al. [69] 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 [17, 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 [38], 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 [7]. Deep learning based methods such as [7] 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 [78, 141]; Vestner et al. [141] 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 chapter, 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 [89] and [141]. In addition, we analyze this measure from a graph theoretic point of view using the statistical model we presented in [48] 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 [92]. 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 chapter is organized a follows. In Section 6.2, 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 6.3. In Section 6.4, we analyze the complexity of the proposed measure. Finally, in Section 6.5, we demonstrate applications of audio-visual localization and eye-fixation prediction.

6.2 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:

$$K_v(n,m) = \exp\left(-\frac{\|\mathbf{v}_n - \mathbf{v}_m\|^2}{\epsilon_v}\right), \tag{6.1}$$

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:

$$\mathbf{M}_v = \mathbf{D}_v^{-1}\mathbf{K}_v, \tag{6.2}$$

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" [79, 135]. 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.

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

6.3 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 [89] and [141], we discuss in Subsection 6.3.1 variants of the proposed measure in the context of kernel density estimation. In Subsection 6.3.2, 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 6.3.3.

6.3.1 From the Perspective of Kernel Density Estimation

The study presented in [141] 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:

$$\arg\max_{\mathbf{P}} \operatorname{Tr}\left\{\mathbf{K}_v \mathbf{K}_w^T \mathbf{P}\right\}, \tag{6.3}$$

where $\mathbf{K}_v, \mathbf{K}_w \in \mathbb{R}^{N\times\tilde{N}}$ are defined similarity to (6.1) 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 [89] 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 minimum mean square error (MMSE) estimator of the data in one modality based on the other. They showed that the corresponding discretized operator is the kernel product:

$$\mathbf{M} = \mathbf{M}_v \mathbf{M}_w^T, \tag{6.4}$$

so that it can replace the kernel $\mathbf{K}_v\mathbf{K}_w^T$ in (6.3) 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 [89], 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 [89] 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 [57]. 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 6.5.

6.3.2 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 (6.1) defines a graph, whose vertices correspond to the N data-points, and the weights of the edges are given by $K_v(n,m) = \exp\left(-\frac{\|\mathbf{v}_n - \mathbf{v}_m\|^2}{\epsilon_v}\right)$ (6.1). 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 [48, 67] 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 [48]. 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:

$$\mathbb{I}_{n,m}^{v} = \left\{ \begin{array}{ll} 1, & \text{w.p. } p_v \\ 0, & \text{otherwise} \end{array} \right\}, \tag{6.5}$$

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 (6.5):

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

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:

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

When the two modalities are uncorrelated, we have:

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

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

have:

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

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.

6.3.3 Online Computation of the Multi-modal Measure of Correspondence

We propose an algorithm for an efficient update of the trace of the kernel product, $\mathrm{Tr}\left\{\mathbf{M}\right\}$, 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 6.3.3, 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. The trace of the kernel product is given by:

$$\mathrm{Tr}\left\{\mathbf{M}\right\} = \sum_{n=1}^{N} D(n,n)K(n,n) \tag{6.6}$$

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

$$\mathrm{Tr}\left\{\mathbf{M}\right\} = \mathrm{Tr}\left\{\mathbf{M}_v\mathbf{M}_w^T\right\} = \mathrm{Tr}\left\{\mathbf{D}_v^{-1}\mathbf{K}_v\left(\mathbf{D}_w^{-1}\mathbf{K}_w\right)^T\right\} = \mathrm{Tr}\left\{\mathbf{D}_v^{-1}\mathbf{K}_v\mathbf{K}_w^T\mathbf{D}_w^{-T}\right\}.$$

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

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

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

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

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

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

Finally, since $\mathbf{D}$ is diagonal, we obtain (6.6). Next we show how to sequentially update (6.6) 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:

$$\mathrm{Tr}\left\{\tilde{\mathbf{M}}\right\} = \sum_{n=1}^{N} \tilde{D}(n,n)\tilde{K}(n,n), \tag{6.7}$$

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 $\tilde{K}(n,n)$ is given by:

$$\tilde{K}(n,n) = \sum_{m=2}^{N+1} K_v(n,m)K_w(n,m), \tag{6.8}$$

so that it is sequentially updated by:

$$\tilde{K}(n,n) = K(n,n) - K_v(n,1)K_w(n,1) + K_v(n,N+1)K_w(n,N+1). \tag{6.9}$$

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

$$\tilde{D}(n,n) = \frac{1}{\tilde{D}_v(n,n)\,\tilde{D}_w(n,n)}, \tag{6.10}$$

where:

$$\tilde{D}_v(n,n) \triangleq \sum_{m=2}^{N+1} K_v(n,m). \tag{6.11}$$

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

$$\tilde{D}_v(n,n) = D_v(n,n) - K_v(n,1) + K_v(n,N+1). \tag{6.12}$$

We summarize the proposed algorithm for the efficient update of the kernel product in Algorithm 6.1.

6.4 Complexity analysis and run-time simulation

We analyze the computational complexity of updating the trace of the kernel product according to Algorithm 6.1. Equation (6.7) 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 (6.10) $2N$ 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 (6.12) by three summations each, which gives a total of $6N$ summations. Finally, computing $\tilde{K}(n,n)$ in (6.9) 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 $O(N)$, and specifically, it requires $10N$ summations and $5N$ multiplications. We further note that in practice, we calculate (6.7),

Algorithm 6.1 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 (6.1)
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 (6.1):
$K_v(n, N+1)$ and $K_w(n, N+1)$, $n \in (2, 3, ..., N)$
5: Update $\tilde{K}(n, n)$ according to (6.9)
6: Update $\tilde{D}_v(n, n)$, $\tilde{D}_w(n, n)$ according to (6.11)
7: Update $\tilde{D}(n, n)$ according to (6.12)
8: Update $\mathrm{Tr}\left\{\tilde{\mathbf{M}}\right\}$ according to (6.6)

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

(6.9), (6.10) and (6.12) simultaneously for $n \in (1, 2, ..., N)$ by writing 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 $O(N^3)$ using naive matrix multiplication methods, and even when efficient algorithms are used, the complexity remains above $O(N^2)$. We relate to this alternative approach as "single modal update" since it was studied in [80] in the single-modal setting.

To demonstrate the run-time efficiency of Algorithm 6.1, 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. 6.1 (left). 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 from. 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 [102]. In addition, we compare the run-time of the proposed approach to an implementation of kernel CCA taken from [1].

We set the number of features in this experiment to $L_v = L_w = 200$ and present the results in Fig. 6.1 (right). 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 [146]. These methods, however, focus on reducing the memory consumption during the processing of large datasets rather than on online processing as in this chapter, and they merely approximate kernel CCA using pre-trained models.

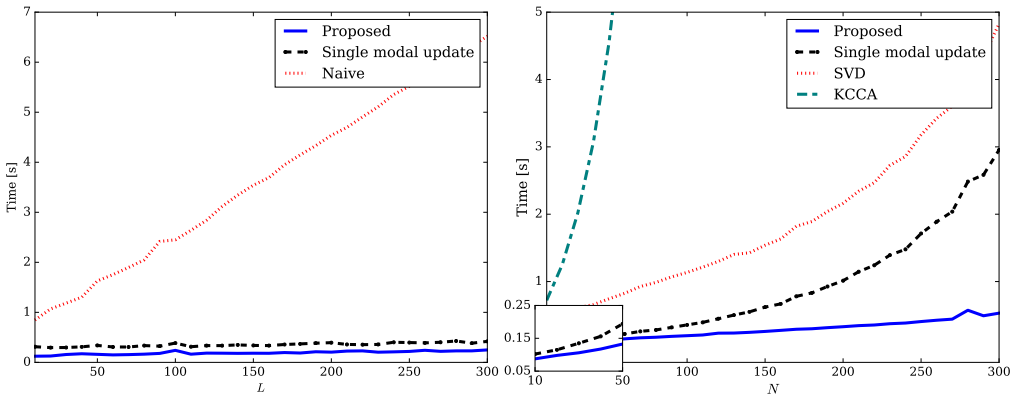


Figure 6.1: Run-time of the algorithms for 1000 new incoming frames averaged over 100 simulations. (left) run-time versus the number of features $L = L_v = L_w$, the batch size is set to $N = 100$. (right) run-time versus batch size in frames N, the number of features is set to $L_v = L_w = 200$.

6.5 Experimental Results

6.5.1 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

[2]link to the videos presented in Figs. 6.2 (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. 6.2 (c) and (d): https://www.youtube.com/watch?v=hx1mjT73xYE, time intervals: $3^{'}:15-4^{'}:45^{''}$ and $4^{'}:55^{''}-6^{'}:25^{''}$

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 [10, 22, 137]. 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.

The audio signal is sampled at 44100 kHz and is processed with time frames of ~ 66 ms with 50% overlap such that the audio and the video signals are aligned. We use 13 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 [85], and we have previously exploited them in [48].

We measure the correspondence between the audio signal and each one of the video streams using the proposed measure in (6.6) based on the product of kernels. We set the kernel bandwidths ϵ_v and ϵ_w according to [67] such that ϵ_v is given by:

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

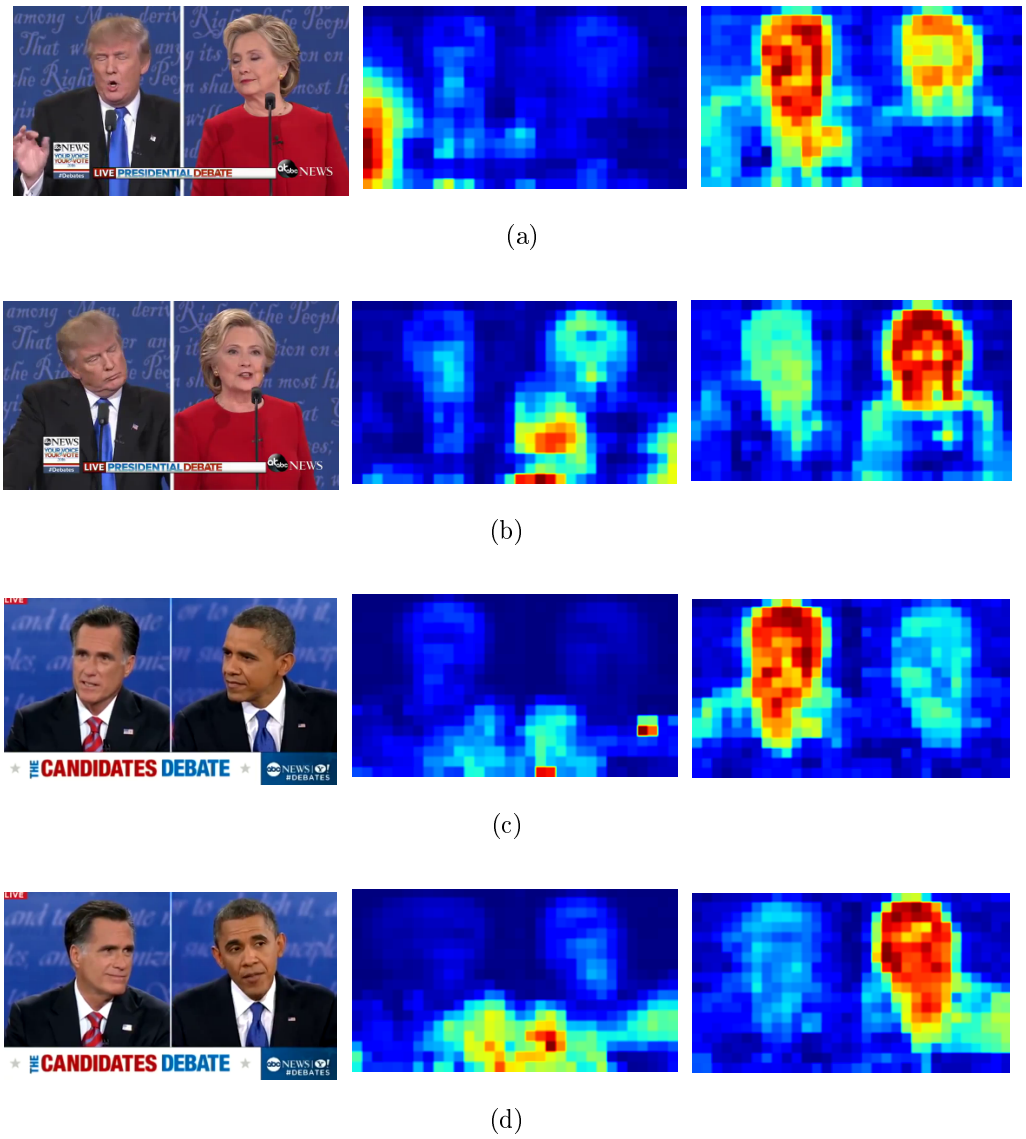
From a graph point of view, each point in the graph is connected when $C = 1$, and C is empirically set to the range of $2 - 3$ to guarantee connectivity of the graph. In [48], we analyzed the selection of the kernel bandwidth in the multi-modal case and showed that the graph which corresponds to the 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. 6.2 (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

[14, 87, 149]. Interestingly, the heat maps of Hillary Clinton and Donald Trump in Figs. 6.2 (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. 6.2 (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. 6.2 (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 [79,135], the proposed measure attenuates sensor specific factors, i.e., the movement of hands, allowing to successfully measure correspondence between the modalities.

6.5.2 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 [92]. 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



(a)

(b)

(c)

(d)

Figure 6.2: 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.

the audio and the video recordings. For more details regarding the dataset, we refer the reader to [92].

We compare the performance of the proposed measure of multi-modal correspondence to the method presented in [92]. 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 signal. The video signal is first divided into super-voxels using a graph-based streaming segmentation method [148]. 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 [92]. 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 [92].

Similarly to [92], we use three common measures to evaluate the prediction of eye fixations. First is the shuffled area under the curve (sAUC), in which 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 shuf-

fled 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 [104]. 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 [92]. 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 [92], 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. 6.3, 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. 6.3 for $N = 25$, a value which we use for comparison to the other methods. Figure 6.3 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 6.1.

In Fig. 6.4, we present eye-fixation predictions obtained by the proposed algorithm in the form of heat maps. In addition to [92], we also compare the proposed method to kernel CCA. Figure 6.4 (a) comprises an example of a video frame of a person playing a piano such that the true eye fixations

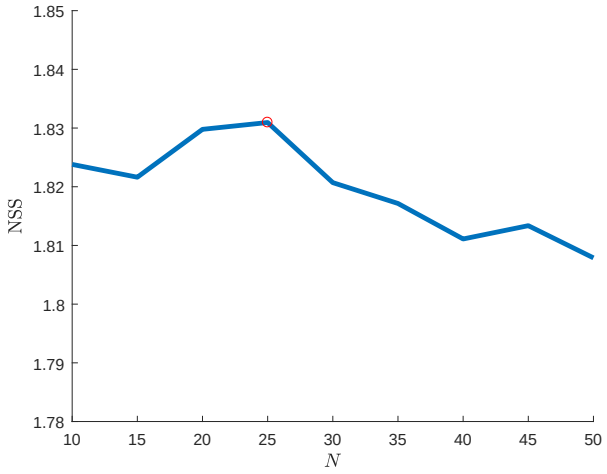


Figure 6.3: The performance of the proposed measure of correspondence for eye-fixations prediction in terms of NSS versus N, the batch size in frames (blue solid line). Best performance obtained for $N = 25$ (red circle).

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 [92], wrongly predicts the walking women in the background. Similarly, both [92] and kernel CCA wrongly predict as salient the movements in the background in Figs. 6.4 (b-d) and the arms movements of the player in Fig. 6.4 (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 6.1. We consider kernel CCA, the empirical HSIC and the method presented in [152] as alternative approaches for measuring correspondence between the audio and video modalities. The method in [152] is based on the use of kernel CCA with multiple kernels and suggests to measure correspondence according to the average distance between

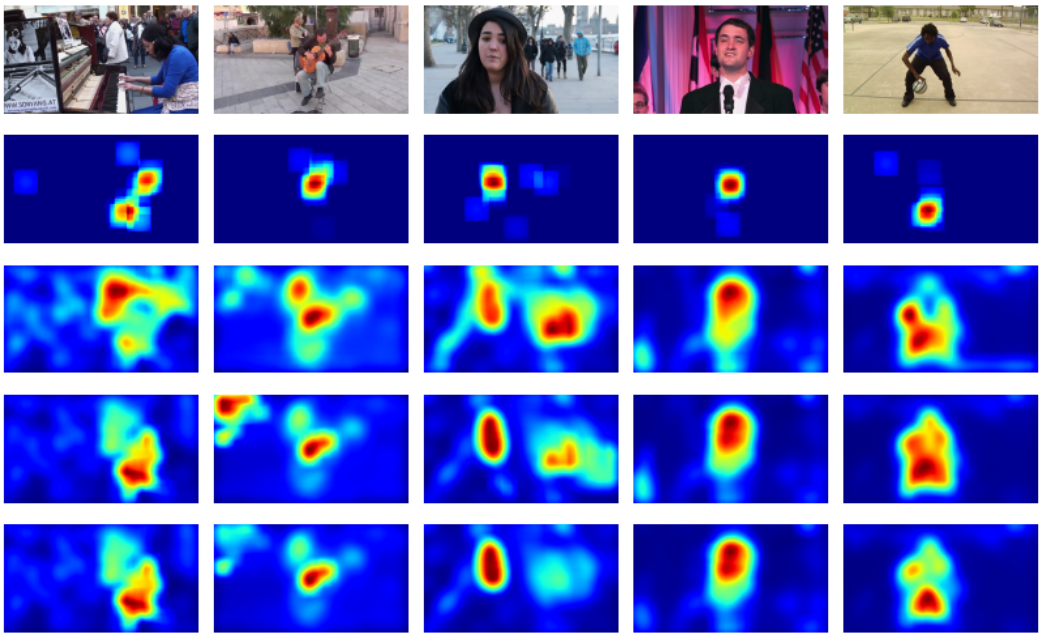
Table 6.1: Comparison of the eye-fixation prediction scores.

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**

audio and video in the space of the kernels. We also compare the proposed method to [64], which is similar to [92] 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 [92] based only on the video signal such that only the spatial and temporal cues are used for the prediction of eye-fixation. The latter approach provides inferior performance, particularly when compared to the proposed method and the method in [92] indicating the significance of the audio signal for eye-fixation prediction. The proposed method provides improved performance compared to the competing approaches.

6.5.3 Discussion

Talmon and Wu provide in [135] 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



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

Figure 6.4: 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 [92], a heat map obtained by kernel CCA, a heat map obtained by the proposed measure of correspondence.

implicitly represents data according to common hidden factors, is empirically supported by Fig. 6.4 such that, for example, background movements are almost completely attenuated in the videos.

Kernel CCA is more sensitive to interferences as demonstrated in Fig. 6.4, 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 [89] 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 6.1 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 [92] 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 [92].

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 chapter. 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 (6.7), (6.9) and (6.11) 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 (6.9) 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. 6.1. 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 [50]. 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. 6.1. 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 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. The frame rate in such a task may be up to 1000 fps as we considered in the single modal setting in [59]. 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.

6.6 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.

Chapter 7

Kernel Method for Voice Activity Detection in the Presence of Transients

Voice activity detection in the presence of transient interferences is a challenging problem since transients are often detected incorrectly as speech by existing detectors. In this chapter, we deviate from traditional approaches and take a geometric standpoint, in which the key element in obtaining an accurate voice activity detection is finding a metric that appropriately distinguishes between speech and transients. For example, speech and transients may often appear similar through the Euclidean distance when represented, e.g., by the Mel-Frequency Cepstral Coefficients, thereby resulting in incorrect speech detection. To address this challenge, we propose to use a metric based on the statistics of the signal in short temporal windows and justify its use by modeling speech and transients by their latent generating variables. These latent variables may be related to physical constraints controlling the generation of the signal, and, as such, they accurately represent the content of the signal – speech or transient. We show that the Euclidean distance between the latent variables is approximated by the proposed metric. Then,

by incorporating this metric into a kernel-based manifold learning method, we devise a measure of voice activity and show it leads to improved detection scores compared with competing detectors. Speech processing, voice activity detection, transient noise, impulse noise, kernel

7.1 Introduction

Signals measured in microphones are often contaminated with various environmental noises and interferences. The environmental conditions pose great challenges in a variety of speech processing tasks, e.g., in speech enhancement [29–32], voice activity detection [25–27,107,108,114,115,123] and dominant speaker identification [142]. Here, we focus on the task of voice activity detection in signals measured in a single microphone, i.e., dividing segments of the signal into speech and non-speech clusters.

To appropriately handle noisy environments, a common approach in the literature is to track the statistics of the signal by recursive averaging in short time intervals [29–32]. It relies on the assumption that the spectrum of the noise slowly varies in time, whereas the spectrum of speech changes quickly. Hence, sudden variations of the spectrum indicate the presence of speech. Although methods based on this statistical approach successfully distinguish speech from quasi-stationary noises, they fail in distinguishing speech from transients, which are abrupt interferences, such as, knocks, keyboard taps and office noise [46,59,129,131]. Since the spectrum of such transients varies in time even quicker than the spectrum of speech, transients are wrongly detected as speech using approaches based on recursive averaging.

To overcome the limitations of existing approaches, recent studies have proposed to model transients according to their geometry [47, 48, 99, 130, 131, 133]. The main assumption in these studies is that transients contain an underlying geometric structure which can be inferred from the signal measurements using manifold learning tools, e.g., those presented in [13, 18,

34,45,110]. In the studies presented in [130,131,133], the geometric structure of transients is captured and is exploited to construct an estimator of their spectrum. In turn, the estimated spectrum is incorporated into a denoising filter and is used for speech enhancement. We emphasize that while these studies deal with the estimation of the spectrum of transients, the present study focuses on the problem of distinguishing them from speech.

In [99], an improved distinction between speech and non-speech frames is obtained by a method based on clustering the noisy signal in a specifically designed low-dimensional domain. More precisely, the method is based on representing time frames of the noisy signal using the MFCCs, and then building a low-dimensional representation of the signal based on local similarities between them. However, the similarities between frames are defined based on the Euclidean distance, which often induces high similarities between speech and transients in standard domains such as the MFCCs and the Short-Time Fourier Transform (STFT) [99, 131]. This results in an incorrect identification of speech and transients, as we demonstrate in this chapter.

To deal with this problem, we use a modified version of the Mahalanobis distance [86], which is constructed from the signal measurements and exploits the statistics of the signal in short temporal windows. We analyze the modified Mahalanobis distance using a model of latent variables; we assume that speech and transients are driven by two independent sets of latent variables controlling their generation and refer to them as *the generating variables*. For example, the generation of the complex speech signal is controlled by the few parameters of the vocal tract [127]. The main idea underlying our approach is that comparing signal frames according to the generating variables gives rise to an accurate detection of the content of the frame, particularly, in terms of speech and transients. The challenge is that these variables are unknown and need to be inferred from the noisy signal. We show that the modified Mahalanobis distance locally approximates the Euclidean distance

in a domain related to the generating variables.

A particular challenge in the problem of voice activity detection is that speech needs to be detected in frames containing both speech and transients. We found in our experiments that transients are often more dominant than speech, i.e., speech frames containing transients tend to be more similar to frames containing only transients, a fact that hampers the detection of speech presence/absence. The dominance of the transients is related to the high variation of their spectrum in time, to their high amplitudes, and to their typical broad bandwidth. We further show that the modified Mahalanobis distance mitigates the problem and reduces the dominance of the transients by implicitly exploiting the respective difference in the rate of variations of speech and transients [53, 59, 131].

By incorporating the modified Mahalanobis distance in a kernel based manifold learning method, we propose an algorithm for voice activity detection. Since this metric approximates the Euclidean distance between the generating variables, the eigenvectors of the kernel provide a parameterization of the signal in terms of the generating variables, which is viewed as the canonical representation of the signal. We show that this canonical representation improves the distinction between speech and transients compared with the representation obtained using the Euclidean distance. In addition, the canonical representation enables us to define a simple measure of voice activity, which outperforms competing detectors.

It is worthwhile noting that classical Voice Activity Detectors (VAD), such as those presented in [25–27, 107, 108, 114, 115, 123], are originally designed to detect speech in the presence of (quasi-) stationary background noise. Based on the assumption that the spectrum of speech rapidly varies compared to the spectrum of (quasi-) stationary background noise, such algorithms detect speech by tracking rapid variations in the spectrum of the noisy signal. In the presence of transients, whose spectrum also rapidly varies over time, such algorithms successfully distinguish the background noise from

both speech and transients, but they cannot distinguish between speech and transients. Consequently, in this chapter, we focus on distinguishing between speech and transients; in practice, a classical VAD may be applied as a pre-processing stage to distinguish time intervals containing only background noise from both speech and transients.

The remainder of the chapter is organized as follows. In Section 7.2, we formulate the problem of voice activity detection. In addition, we present a metric based on the statistics of the signal in short temporal windows, and to justify its use, we propose a model of latent generating variables. Based on this model, we show in Section 7.3 that the metric reduces the effect of transients. Using this metric, an algorithm for voice activity detection is introduced in Section 7.4, and experimental results demonstrating the superior performance of the proposed algorithm are presented in Section 7.5.

7.2 Problem Formulation

7.2.1 The Problem of Voice Activity Detection

Consider a speech signal obtained in a single microphone in the presence of transients and processed in frames. Let $\mathbf{y}_n \in \mathbb{R}^L$ be a feature representation of frame n; in particular, we use the MFCCs such that L is the number of coefficients. The MFCCs are widely used features for speech representation based on the perceptually meaningful Mel-frequency scale [85]. They represent the spectrum of the signal in a compact form and they were previously exploited both for speech recognition [41, 58] and for voice activity detection [70]. We consider a setup where a sequence of N frames $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N$ is available in advance. Assume that the sequence comprises frames where speech is present and frames where it is absent, and let $\mathcal{H}_1^x$ and $\mathcal{H}_0^x$ be two hypotheses representing the presence and the absence of speech, respectively, where x denotes speech. Based on the hypotheses, we define a speech indi-

cator for frame n, which is denoted by $\mathbb{1}_n^x$ and is given by:

$$\mathbb{1}_n^x = \left\{ \begin{array}{ll} 1; & n \in \mathcal{H}_1^x \\ 0; & n \in \mathcal{H}_o^x \end{array} \right\}. \tag{7.1}$$

The objective in this study is to estimate the speech indicator, i.e., to cluster the sequence according to the two hypotheses.

Similarly to the hypotheses $\mathcal{H}_1^x$ and $\mathcal{H}_0^x$, let $\mathcal{H}_1^t$ and $\mathcal{H}_0^t$ be hypotheses of the presence and the absence of transients, respectively, where t denotes transients. We note that frames containing both speech and transients, for which both hypotheses $\mathcal{H}_1^x$ and $\mathcal{H}_1^t$ hold, are considered as speech frames for the purpose of voice activity detection. Nevertheless, transients are typically more dominant than speech, e.g., due to higher amplitudes and broader bandwidth. Accordingly, the MFCCs of frames containing both speech and transients often appear similar to the MFCCs of frames containing only transients, and, as a result, they are often wrongly identified as we show in Section 7.5. One approach for improving the clustering is to design features, for which the Euclidean distance better distinguishes between the content of the frames of the signal. In this study, we take a different approach and propose to use a different metric instead of the Euclidean distance. Specifically, we propose to measure distances between frames of the signal using a modified Mahalanobis distance, proposed in [86], which is given by:

$$\|\mathbf{y}_n - \mathbf{y}_m\|_M^2 \triangleq \frac{1}{2} \left(\mathbf{y}_n - \mathbf{y}_m\right)^T \left(\mathbf{C}_n^{-1} + \mathbf{C}_m^{-1}\right) \left(\mathbf{y}_n - \mathbf{y}_m\right), \tag{7.2}$$

where $\mathbf{C}_n \in \mathbb{R}^{L \times L}$ and $\mathbf{C}_m \in \mathbb{R}^{L \times L}$ are the covariance matrices of $\mathbf{y}_n$ and $\mathbf{y}_m$, respectively. The covariance matrices are assumed to be known throughout this section and throughout Section 7.3; in Section 7.5, we describe their estimation from a short time window of samples. The modified Mahalanobis distance was previously presented in [119] for the purpose of solving the problem of non-linear independent component analysis, in which the assumption

is that the observable signal is generated by independent latent stochastic dynamical processes. However, these processes are assumed to smoothly evolve in time, i.e., the current state of the process is correlated with previous states. Therefore, such processes cannot properly model transitions between speech presence and absence. Hence, to justify the use of the modified Mahalanobis distance in (7.2) for voice activity detection, we propose in Section 7.2.2 to model the noisy signal using latent variables controlling its generation. By assuming a simplifying statistical model for the generating variables, we show in Section 7.3 that the modified Mahalanobis distance approximates a weighted Euclidean distance between the variables, which properly respects the content of the noisy signal.

7.2.2 The Model of The Generating Variables

The generation of many signals can be associated with a small set of physical constraints controlling their production. For example, the generation of speech is controlled by the position of the vocal tract and by the movement of lips, jaw, and tongue [127]. Here, we assume that the measured signal is modeled by two sets of unknown latent variables associated with the generation of speech and the transients. Let $\boldsymbol{\theta}_n^x \in \mathbb{R}^{d^x}$ and $\boldsymbol{\theta}_n^t \in \mathbb{R}^{d^t}$ be two vectors of generating variables underlying the speech signal and the transients in frame n, where d^x and d^t are the number of the variables, respectively. The vector of all generating variables at time frame n is denoted by $\boldsymbol{\theta}_n \in \mathbb{R}^d$, where $d \triangleq d^x + d^t$, and is given by:

$$\boldsymbol{\theta}_n = \left[\left(\boldsymbol{\theta}_n^x\right)^T, \left(\boldsymbol{\theta}_n^t\right)^T \right]^T, \tag{7.3}$$

where T denotes transpose. The generating variables are assumed hidden, i.e., $\boldsymbol{\theta}_n$ in (7.3) is not directly measured by the microphone. For example, a variable that is related to the movement of the lips during the production of speech cannot be directly captured in the microphone.

We assume that the relationship between the observable signal $\mathbf{y}_n$ and the vector of the generating variables $\boldsymbol{\theta}_n$ is given by an unknown non-linear transformation $f : \mathbb{R}^d \mapsto \mathbb{R}^L$, such that:

$$\mathbf{y}_n = f\left(\boldsymbol{\theta}_n\right). \tag{7.4}$$

If we had access to the generating variables, then voice activity detection would become trivial since one may ignore the variables of the transients, $\boldsymbol{\theta}_n^t$, and detect speech merely from the variables of speech, $\boldsymbol{\theta}_n^x$. However, the generating variables are not directly accessible and revealing them is challenging due to their unknown non-linear mapping f in (7.4) to the observable domain. Still, in the sequel, we assume a simplified model for the generating variables and the non-linear transformation f in (7.4), and based on this model, we show in Section 7.3 that the modified Mahalanobis distance in (7.2) approximates weighted Euclidean distances between frames in the domain of the generating variables. Specifically, we will show that the proposed metric reduces the effect of transients, thereby allowing improved distinction of frames containing both speech and transients from frames containing merely transients. We emphasize that in practice the generating variables are not directly estimated, but used for the analysis of the modified Mahalanobis distance in (7.2).

We first assume that the generating variables are statistically independent such that $\boldsymbol{\theta}_n$ has a diagonal covariance matrix. The variables of speech $\boldsymbol{\theta}_n^x$ and the variables of transients $\boldsymbol{\theta}_n^t$ are assumed independent since they are related to two independent phenomena - speech and transients. The independence between each of the variables of (say) speech, i.e., between the entries of $\boldsymbol{\theta}_n^x$, may be associated with a lack of correlation between the corresponding physical constraints. For example, the pronunciation of different parts of speech, e.g., different phonemes, is based on different combinations of the position of the vocal tract and the movement of lips, jaw, and tongue. We note that the independence between variables is a common assumption found

in the literature for the analysis of latent models [11, 119, 138]. For example, in [138], the authors suggest a model of latent IID variables to provide a probabilistic interpretation of the classical Principal Component Analysis (PCA).

To encode the dominance of the transients, we assume that the generating variables of the transients have larger variances than the variables of speech. Specifically, to keep the statistical model simple, we assume that under hypotheses $\mathcal{H}_1^x$ and $\mathcal{H}_1^t$, the entries of $\boldsymbol{\theta}_n^x$ and $\boldsymbol{\theta}_n^t$ are IID, with zero mean, and $\sigma_x^2 > 0$ and $\sigma_t^2 > 0$ variances, respectively. We assume that:

$$\sigma_t^2 = r^2 \sigma_x^2, \tag{7.5}$$

where $r^2 > 1$ is a constant factor encoding the dominance of the transients, such that a larger r implies more dominant transients. The parameter r may be seen as related to the ratio between transients and speech. Typically, even when the transients and speech are normalized to the same maximal value, transients, due to their short duration in time, are more dominant than speech. We note that in order to show in Section 7.3 the link between the modified Mahalanobis distance and the generating variables, we do not assume specific distributions of the generating variables and they do not have to be identically distributed. In particular, the variances of the generating variables of (say) speech, i.e., the entries of $\boldsymbol{\theta}_n^x$ do not necessarily equal to the same value σ_x^2, but they are only assumed to have larger variances than the variables of speech. In Section 7.3, we show that the modified Mahalanobis distance approximates weighted distances between the generating variables such that the weights reduce the effect of the more dominant variables, which are the transients, by assumption. Namely, we link the variances of the variables of speech and transients by a single parameter r only for the sake of simplicity. In addition, the mean value of the generating variables is set to zero merely for simplicity and it is not used explicitly in this study. Under the hypotheses $\mathcal{H}_0^x$ and $\mathcal{H}_0^t$, we simply assume that the generating variables

of speech and transients equal zero, respectively. Thus, in the presence of speech only, for example, the observable signal $\mathbf{y}_n$ is related only to the generating variables of speech and not to those of the transients.

For the approximation in Section 7.3 showing the relation between the modified Mahalanobis distance and the generating variables, we consider the inverse of the function f in (7.4). However, we consider only frames located within a local neighborhood such that the (Euclidean) distance between them is smaller than a certain value. In such neighborhoods, we assume that the function f in (7.4) is smooth and *locally* invertible. Note that this assumption is significantly less restrictive than assuming a globally invertible function. In this context, we further note that in Section 7.4 we take a data-driven approach to obtain a representation of the noisy signal based on the generating variables, by exploiting the Mahalanobis distances between the frames of the measured signal. Accordingly, the assumption that the function f in (7.4) is locally invertible is not strictly imposed, i.e., if it does not hold in practice, the obtained representation may be seen as the best fit of the model of the generating variables to the measured signal.

To facilitate the model of the generating variables, the presence of a (quasi-) stationary noise is not considered in this chapter. In practice, a classical speech enhancement algorithm, e.g., the one presented in [31], may be used as a preprocessing stage to attenuate stationary noise. Such an algorithm is based on the assumption that the spectrum of the speech signal rapidly varies over time compared to the spectrum of a (quasi-) stationary noise. Hence, the stationary noise is estimated (and then attenuated) by tracking the small variations of the spectrum of the noisy signal. Since the spectrum of transients also rapidly varies over time, it is "seen" by such a speech enhancement algorithm as speech. As a result, the speech enhancement algorithm attenuates only the stationary noise while preserving speech and the transients. Accordingly, we assume that frames which do not contain speech nor transients, i.e., silent frames, are known in advance focusing on

the more challenging problem of distinguishing speech from transients. Silent frames are successfully identified even in the presence of stationary noise by classical voice activity detectors, e.g., those presented in [108, 123].

7.3 Modified Mahalanobis Distance

In this section we show that the modified Mahalanobis distance (7.2) approximates the following distance:

$$\|\mathbf{y}_n - \mathbf{y}_m\|_M^2 = \tfrac{1}{\sigma_x^2}\left(\|\boldsymbol{\theta}_n^x - \boldsymbol{\theta}_m^x\|^2 + \tfrac{1}{r^2}\left\|\boldsymbol{\theta}_n^t - \boldsymbol{\theta}_m^t\right\|^2\right) + O\left(\|\mathbf{y}_n - \mathbf{y}_m\|^4\right), \tag{7.6}$$

which consists of a weighted sum of the Euclidean distances between the generating variables.

As we observed in our experiments, the main challenge in obtaining a successful clustering arises from the fact that speech components may be similar to transient components. Consider, for example, two frames, $\mathbf{y}_n$ and $\mathbf{y}_m$, one consists of only speech and the other consists of only transients. Often, small Euclidean distances are obtained between the MFCCs of such pairs of frames; as a result, these frames are not properly associated with different clusters as we demonstrate in Section 7.5 [99, 131]. However, speech and transients are assumed to have different generating variables. As a result, the Mahalanobis distance (7.6) between these frames is given by

$$\|\mathbf{y}_n - \mathbf{y}_m\|_M^2 \approx \frac{1}{\sigma_x^2}\left(\|\boldsymbol{\theta}_n^x\|^2 + \frac{1}{r^2}\left\|\boldsymbol{\theta}_m^t\right\|^2\right). \tag{7.7}$$

Hence, the distance between these two frames is given according to the squared norms of $\boldsymbol{\theta}_n^x$ and $\boldsymbol{\theta}_m^t$ conveying the different content of the frames, in contrast to the Euclidean distance. Assuming for simplicity that $\sigma_x = 1$, this example demonstrates that the content of a frame is better represented by the Mahalanobis distance, which approximates the Euclidean distance be-

tween the generating variables, i.e., a small Mahalanobis distance between frames truly indicates that they comprise a similar content.

Another property of the Mahalanobis distance (7.6) stems from the re-scaling of the Euclidean distance between the generating variables of the transients by a factor of r^2. Since transient components are often more dominant than speech components due to their typical abrupt nature and large amplitudes, frames containing both speech and transients tend to be labeled as "transient" frames, i.e., $\mathcal{H}_1^t$, by typical clustering algorithms. This poses a problem for voice activity detection, where the speech presence is required to dominate the clustering. The Mahalanobis distance (7.7) mitigates the dominance of transients by reducing the weight of the Euclidean distance between their generating variables by a factor of $r^2 > 1$, thereby allowing for the design of a voice activity detector in which transients are less dominant, and frames tend more to be labeled according to their speech presence and absence, as demonstrated in Section 7.5.

We note that the approximation in (7.6) holds only for short distances, where the error term $\|\mathbf{y}_n - \mathbf{y}_m\|^4$ is negligible. In Section 7.4 we show how to obtain a global representation of the generating variables by incorporating this metric in a kernel-based manifold learning method.

To derive (7.6), we follow [53]. Consider the re-scaled vectors $\boldsymbol{\psi}_n^x \in \mathbb{R}^{d^x}$ and $\boldsymbol{\psi}_n^t \in \mathbb{R}^{d^t}$ defined by:

$$\boldsymbol{\psi}_n^x = \frac{\boldsymbol{\theta}_n^x}{\sigma_x} \tag{7.8}$$

$$\boldsymbol{\psi}_n^t = \frac{\boldsymbol{\theta}_n^t}{\sigma_t} \tag{7.9}$$

such that the entries of the vectors have unit variances. In addition, let $\boldsymbol{\psi}_n \in \mathbb{R}^d$ denote a vector consisting of all the re-scaled variables in the nth frame:

$$\boldsymbol{\psi}_n = \left[\left(\boldsymbol{\psi}_n^x\right)^T, \left(\boldsymbol{\psi}_n^t\right)^T \right]^T, \tag{7.10}$$

and let $h : \mathbb{R}^d \mapsto \mathbb{R}^L$ denote the corresponding nonlinear function that maps

the re-scaled variables to the observable signal:

$$\mathbf{y}_n = h\left(\boldsymbol{\psi}_n\right). \tag{7.11}$$

The function h is locally invertible since we assume that the function f in (7.4) is locally invertible; consequently, let $g : \mathbb{R}^L \mapsto \mathbb{R}^d$ be an inverse map of h, i.e., $\boldsymbol{\psi}_n = g\left(\mathbf{y}_n\right)$. Note that for simplicity we follow [119], and, for all points $\mathbf{y}$ considered throughout the chapter, we denote by $g\left(\mathbf{y}\right)$ the local inverse map of the function h even though the function h is assumed invertible only locally.

Singer et al. derived (7.6) in [119, 120] by using the Taylor expansions of $\boldsymbol{\psi}_n = g\left(\mathbf{y}_n\right)$ and $\boldsymbol{\psi}_m = g\left(\mathbf{y}_m\right)$ at $\mathbf{y}_m$ and $\mathbf{y}_n$, respectively, relying on the symmetry of the expansions. However, in our case, two frames $\mathbf{y}_n$ and $\mathbf{y}_m$ may consist of different signals, e.g, $\mathbf{y}_n$ may consist of only speech and $\mathbf{y}_m$ may consist of only transients, thereby breaking the symmetry between the Taylor expansions of $\boldsymbol{\psi}_n$ and $\boldsymbol{\psi}_m$.

To overcome this problem, we consider the middle point $\mathbf{y}_p$ between $\mathbf{y}_n$ and $\mathbf{y}_m$:

$$\mathbf{y}_p = \frac{\mathbf{y}_n + \mathbf{y}_m}{2}, \tag{7.12}$$

which does not necessarily exist in practice, but is used here merely as a reference point for the derivation. The mid-point relaxes the symmetry assumption since it contains speech or transients if they are present in one of the frames $\mathbf{y}_n$ or $\mathbf{y}_m$.

First, we focus on the hypothesis that both speech and transients are present, and then extend the derivation to all other possible hypotheses. Specifically, $\mathbf{y}_n$ and $\mathbf{y}_m$ are assumed to contain both speech and transients, and hence, so is the mid-point $\mathbf{y}_p$.

Kushnir et al. have shown in [75] that using a second order Taylor expansions of $\boldsymbol{\psi}_n$ and $\boldsymbol{\psi}_m$ at the mid-point, the Euclidean distance between

the two points is given by:

$$\|\boldsymbol{\psi}_n - \boldsymbol{\psi}_m\|^2 = (\mathbf{y}_n - \mathbf{y}_m)^T \boldsymbol{\Lambda}^{-1}(\mathbf{y}_p)(\mathbf{y}_n - \mathbf{y}_m) + O\left(\|\mathbf{y}_n - \mathbf{y}_m\|^4\right), \tag{7.13}$$

where $\boldsymbol{\Lambda}^{-1}(\mathbf{y}_p)$ is a pseudo-inverse of $\boldsymbol{\Lambda}(\boldsymbol{\psi}_p) \triangleq \mathbf{J}\mathbf{J}^T(\boldsymbol{\psi}_p) \in \mathbb{R}^{L\times L}$, and $\mathbf{J}(\boldsymbol{\psi}_p) \in \mathbb{R}^{L\times d}$ is the Jacobian of the function g at the mid-point. The approximation to the fourth order in (7.13) holds due to the symmetry of the Taylor expansions of $\boldsymbol{\psi}_n$ and $\boldsymbol{\psi}_m$ at the mid-point under our hypothesis; for the sake of completeness, the derivation of (7.13) is given in Appendix I. In Appendix II, we further show that the term $\boldsymbol{\Lambda}^{-1}(\mathbf{y}_p)$ in (7.13) can be replaced by the term $\frac{1}{2}\boldsymbol{\Lambda}^{-1}(\mathbf{y}_n) + \frac{1}{2}\boldsymbol{\Lambda}^{-1}(\mathbf{y}_m)$; this result is obtained by the Taylor expansion of $\boldsymbol{\Lambda}^{-1}(\mathbf{y}_n)$ and $\boldsymbol{\Lambda}^{-1}(\mathbf{y}_m)$ to the first order at the mid-point. Consequently, we have:

$$\|\boldsymbol{\psi}_n - \boldsymbol{\psi}_m\|^2 = \tfrac{1}{2}(\mathbf{y}_n - \mathbf{y}_m)^T \left(\boldsymbol{\Lambda}^{-1}(\mathbf{y}_n) + \boldsymbol{\Lambda}^{-1}(\mathbf{y}_m)\right)(\mathbf{y}_n - \mathbf{y}_m) \tag{7.14}$$
$$+O\left(\|\mathbf{y}_n - \mathbf{y}_m\|^4\right).$$

where the mid-point, which may not exist in practice, does not appear. Yet, the Jacobian matrices at $\mathbf{y}_n$ and $\mathbf{y}_m$ in (7.14) are unknown. In Appendix III, we show that these Jacobian matrices can be estimated from the signal at hand based on local temporal statistics. Specifically, we show that the terms $\boldsymbol{\Lambda}^{-1}(\mathbf{y}_n)$ and $\boldsymbol{\Lambda}^{-1}(\mathbf{y}_m)$ in (7.14) are equivalent to the inverse of the local covariance matrices $\mathbf{C}_n^{-1}$ and $\mathbf{C}_m^{-1}$, respectively. Thus, using the definition of the modified Mahalanobis distance (7.2), we obtain:

$$\|\boldsymbol{\psi}_n - \boldsymbol{\psi}_m\|^2 = \|\mathbf{y}_n - \mathbf{y}_m\|_M^2 + O\left(\|\mathbf{y}_n - \mathbf{y}_m\|^4\right). \tag{7.15}$$

Reordering (7.15) yields:

$$\|\mathbf{y}_n - \mathbf{y}_m\|_M^2 = \|\boldsymbol{\psi}_n^x - \boldsymbol{\psi}_m^x\|^2 + \left\|\boldsymbol{\psi}_n^t - \boldsymbol{\psi}_m^t\right\|^2 + O\left(\|\mathbf{y}_n - \mathbf{y}_m\|^4\right), \tag{7.16}$$

and substituting the re-scaled variables, $\boldsymbol{\psi}_n^x$ and $\boldsymbol{\psi}_n^t$, by the generating variables, $\boldsymbol{\theta}_n^x$ and $\boldsymbol{\theta}_n^t$, respectively, leads to (7.6).

Thus far, the derivation of (7.6) was made under the assumption that hypotheses $\mathcal{H}_1^x$ and $\mathcal{H}_1^t$ hold for both frames $\mathbf{y}_n$ and $\mathbf{y}_m$; in Appendix IV, we derive (7.6) under the other hypotheses. We note that the limitation of the result in (7.6) lies in the assumption that the covariance matrix $\mathbf{C}_n$ is invertible [75,119,134]. In practice, when the dimension of the generating variables, d, is smaller than the dimension of the observable signal, L, the covariance matrix is not invertible, and, in this case, a pseudo-inverse should be used; we further discuss the estimation of the covariance matrices in Section 7.5.

7.4 Canonical Representation Through Diffusion Maps for Voice Activity Detection

The metric we present in (7.2) approximates the Euclidean distance between the (re-scaled) generating variables; however, the approximation holds only for short distances, where the factor $O\left(||\mathbf{y}(n)-\mathbf{y}(m)||^4\right)$ in (7.6) is negligible. Therefore, the proposed metric cannot be directly incorporated in typical clustering or classification methods such as support vector machines (SVM). To overcome this limitation, we use a kernel-based geometric method, termed *diffusion maps*, with a Gaussian kernel which "sees" only local distances between frames [34]. Diffusion maps integrates all local distances into a global parameterization respecting the local distances; since the local distances are based on the generating variables, this global parameterization represents the generating variables and can be viewed as the canonical representation of the signal.

Let $k\left(\mathbf{y}_n,\mathbf{y}_m\right)$ be a similarity kernel between frames $\mathbf{y}_n$ and $\mathbf{y}_m$, given by:

$$k\left(\mathbf{y}_n,\mathbf{y}_m\right)=e^{-\frac{\|\mathbf{y}_n-\mathbf{y}_m\|_M^2}{\varepsilon}}, \tag{7.17}$$

where ε is a scaling parameter. Short distances between frame $\mathbf{y}_n$ and frame $\mathbf{y}_m$ provide high values of the kernel, whereas distances much greater than the scaling parameter ε are negligible. In practice, we set the parameter ε according to [67]; since for distances smaller than ε the approximation in (7.6) holds, the proposed kernel measures local similarities between frames according to the (re-scaled) generating variables. Using the kernel in (7.17), we construct an affinity matrix $\mathbf{K} \in \mathbb{R}^{N\times N}$ such that its (n,m)th entry, denoted by $K_{n,m}$, represents the similarity between frame $\mathbf{y}_n$ and frame $\mathbf{y}_m$:

$$K_{n,m} = k\left(\mathbf{y}_n, \mathbf{y}_m\right). \tag{7.18}$$

The affinity matrix $\mathbf{K}$ defines a weighted symmetric graph such that the frames $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N$ are the nodes of the graph and the edge between frame $\mathbf{y}_n$ and frame $\mathbf{y}_m$ is given by $K_{n,m}$. We define a Markov chain on the graph by normalizing the kernel [34]:

$$\mathbf{M} = \mathbf{D}^{-1}\mathbf{K}, \tag{7.19}$$

where $\mathbf{D} \in \mathbb{R}^{N\times N}$ is a diagonal matrix with $D_{m,m} = \sum_n K_{m,n}$. Namely, $\mathbf{M} \in \mathbb{R}^{N\times N}$ is a row stochastic Markov matrix whose rows sum to one. Then, we apply the eigenvalue decomposition to $\mathbf{M}$ yielding eigenvalues $1 = \lambda_0 > \lambda_1 > ... > \lambda_{N-1} \in \mathbb{R}$ corresponding to eigenvectors $\boldsymbol{\phi}_0, \boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_{N-1} \in \mathbb{R}^N$ [34]. Due to the row normalization, the leading eigenvalue λ_0 equals one, and the leading eigenvector $\boldsymbol{\phi}_0$ is an all ones vector that we ignore since it does not contain information. We use the eigenvectors to form a global parameterization of the signal. Specifically, we construct a matrix $\boldsymbol{\Phi} \in \mathbb{R}^{N\times J}$ using $J < N$ eigenvectors corresponding to the J largest eigenvalues:

$$\boldsymbol{\Phi} \equiv \left[\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_J\right], \tag{7.20}$$

where the nth row of the matrix is the parameterization of frame $\mathbf{y}_n$. This

parameterization respects the local affinities between the generating variables and is independent of the mapping function f. Therefore we view it as the canonical representation of the signal. In [99], similarly to the present study, the eigenvectors of a kernel function are used to construct a low-dimensional representation of the signal. The representation is exploited for voice activity detection in a supervised learning framework. Specifically, a measure of voice activity is constructed in the low-dimensional domain using a training set comprising marked speech and transients segments. In this study, we obtain improved clustering between speech and transients using the proposed kernel as we demonstrate in Section 7.5. Hence, we take an *unsupervised* approach and propose to use only the leading (non-trivial) eigenvector, $\boldsymbol{\phi}_1$, as a measure of voice activity, i.e., we set $J = 1$ in (7.20). We emphasize that the eigenvector $\boldsymbol{\phi}_1$ is of length N, as the number of frames in the sequence, and each of its coordinates describe a frame. Specifically, we estimate the speech indicator of frame n in (7.1) by comparing the nth entry of $\boldsymbol{\phi}_1$, which we denote by $\phi_1(n)$, to a threshold, such that values above the threshold indicate voice activity:

$$\hat{\mathbb{1}}_n^x = \left\{ \begin{array}{ll} 1; & \phi_1(n) > \tau \\ 0; & \text{otherwise} \end{array} \right\}, \tag{7.21}$$

where τ is the threshold value. The threshold value may control the trade-off between correct detection and false alarm rates, and, in particular, setting the threshold value to zero may provide a good distinction between speech and non-speech frames as we will show in Section 7.5. We note here that the leading eigenvector, $\boldsymbol{\phi}_1$, solves the well-known normalized cut problem presented in [113] and is widely used for clustering. The main difference with respect to previous studies is that in this study, the use of the modified Mahalanobis distance gives rise to the clustering of the signal according to the generating variables. In addition, we use the leading eigenvector as a *continuous measure* of voice activity in contrast to binary labeling. We will show

in Section 7.5 that the leading eigenvector successfully distinguishes between speech and transients and provides improved detection scores compared to competing detectors. The proposed algorithm for voice activity detection is summarized in Algorithm 7.1.

Algorithm 7.1 Voice activity detection

1: Calculate the MFCCs of the noisy signal $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N$ and estimate the corresponding covariance matrices $\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_N$
2: Calculate the affinity kernel $\mathbf{K}$ based on the modified Mahalanobis distance according to (7.2), (7.17) and (7.18)
3: Calculate $\mathbf{M}$ according to (7.19)
4: Obtain the leading eigenvector $\boldsymbol{\phi}_1$
5: **for** $n = 1 : N$ **do**
6: **if** $\phi_1(n) > \tau$ **then**
7: $\hat{\mathbb{1}}_n^x = 1$
8: **else**
9: $\hat{\mathbb{1}}_n^x = 0$
10: **end if**
11: **end for**

7.5 Experimental Results

7.5.1 Implementation

To evaluate the performance of the proposed approach, we use speech and transient signals taken from the TIMIT database [56] and an online free corpus [3], respectively. The signals are sampled at 16 kHz, and are processed in frames of 512 samples with 50 percent overlap. We use 40 speech utterances of different speakers and construct 20 sequences, $20-30$ s long, by raffling 5 random utterances for each sequence. The transients are synthetically added to the speech sequences, and they are normalized to have the same maximal values. This type of normalization was previously used in [99, 131, 133], and we find it more convenient than for example, normalizing the transients

according to their energy, which often has small values due to the short duration of the transients.

The proposed metric in (7.2) requires the estimation of local covariance matrices for each frame of the signal; one approach for their estimation is to use the sample covariance, as was suggested in [109]:

$$\hat{\mathbf{C}}_n = \frac{1}{2R+1} \sum_{i=-R}^{R} \left(\mathbf{y}_{n+i} - \hat{\boldsymbol{\mu}}_n\right) \left(\mathbf{y}_{n+i} - \hat{\boldsymbol{\mu}}_n\right)^T ,$$

where $\mathbf{y}_{n-R}, \mathbf{y}_{n-R+1}, ..., \mathbf{y}_{n+R}$ are consecutive frames at a small temporal neighborhood of frame $\mathbf{y}_n$, and $\hat{\boldsymbol{\mu}}_n = \frac{1}{2R+1} \sum_{i=-R}^{R} \mathbf{y}_{n+i}$ is the sample mean. In our experiments, we set R to 15 and similarly to the finding in [109], we empirically found that a good distinction between speech and transients is obtained using a very small temporal neighborhood with a high overlap between the consecutive frames. However, the use of highly overlapping frames significantly increases the computational cost of the algorithm. Hence, in this study we assume that entries of the observable signal are uncorrelated such that the covariance matrix is diagonal. Accordingly, we estimate the variance of each entry of the observable signal using recursive averaging of the spectrum of the signal, similarly to the method presented in [29]. In this approach, we exploit the entire signal including the silent frames since the variances of speech and the transients are estimated according to variations of the spectrum of the noisy signal with respect to the spectrum estimated in the silent frames. Recall that when the dimension of the generating variables, d, is smaller than the dimension of the observable signal L, a pseudo-inverse is used for the estimation of the inverse of the covariance matrix in (7.2). We empirically found that applying a pseudo-inverse using three entries of the observable signal with the highest variances provide improved distinction between speech and transients. This finding heuristically implies that the signal is controlled by three generating variables. In our experiments, the estimation of the covariance matrices based on recursive averaging of the

spectrum of the signal provides better detection scores compared to the use of the sample covariance, and it is the one used in the simulations in this section. The estimation of the covariance matrix will be further addressed in a future study.

The proposed representation is obtained according to (7.17)-(7.20) in a batch manner since all N frames of the sequence are required in advance to calculate the affinity matrix in (7.18). Still, the proposed algorithm may be implemented in an online manner, e.g., by constructing the canonical representation of the signal using a calibration set, given in advance without labels. Then, the eigenvectors of the kernel may be extended to new incoming frames, e.g., using the Nyström method [55]. In this context we note that to reduce the computational cost of the affinity matrix calculation in (7.18), we exploit a non-symmetric kernel in (7.17), and address the reader to [75] for more implementation details. We empirically found that it provides better detection scores compared to calculating the symmetric kernel.

7.5.2 Voice Activity Detection

The proposed representation of a speech signal, contaminated with a door-knocks transient, is illustrated in Fig. 7.1 (right), and is compared to the representation obtained using the Euclidean distance instead of the Mahalanobis distance in Fig. 7.1 (left). In both figures, we present a scatter plot of the first two eigenvectors of the affinity kernel such that each point represents a time frame. The points are marked according to the hypotheses $\mathcal{H}_1^x$ and $\mathcal{H}_1^t$ using the labels of the ground truth: frames for which only one of the hypotheses, $\mathcal{H}_1^x$ or $\mathcal{H}_1^t$, holds are marked with red squares and green stars, respectively, and those for which both hypotheses $\mathcal{H}_1^x$ and $\mathcal{H}_1^t$ hold are marked with blue circles. It can be seen in Fig. 7.1 (left) that the representation obtained based on the Euclidean distance only partially distinguishes between speech and non-speech frames. In particular, since transients are often more dominant than speech, many frames containing both speech and

transients are represented as similar to frames containing only transients. In contrast, the representation obtained based on the proposed metric, illustrated in Fig. 7.1 (right), provides improved clustering between speech and transients. In particular, frames containing both speech and transients tend to be more similar to speech frames than to transients.

The representation obtained from the noisy signal using the proposed metric allows us to devise a measure of voice activity in an unsupervised manner based on the first eigenvector. Specifically, we can estimate the speech indicator for voice activity in (7.1) by comparing the first eigenvector to a threshold such that values above the threshold indicate voice activity. Specifically, setting the threshold value to zero may provide a good distinction between speech and non-speech frames. At this point we note that the eigenvectors are obtained by the eigenvalue decomposition with arbitrary signs. Therefore, the sign of the first eigenvector has to be set such that the speech cluster corresponds to its high values. In this study, we assume that the correct sign of the eigenvector is known. In practice, the sign of the eigenvector may be set according to the temporal variability of the signal, such that the cluster of transients is assumed to comprise segments of the signal with higher variability rates over time [59]. In addition, we note that although in this study we only use a single eigenvector, more eigenvectors may be used for voice activity detection. For example, in the studies presented in [47,99], several eigenvectors are used as a low dimensional representation and they are incorporated in a supervised learning framework. However, these studies consider a different problem setup, where the type of transients is known in advance and that they are available in a training set. A different heuristic approach, which does not require a training set, is using a deterministic combination between the eigenvectors, e.g., the sum of the first two eigenvectors; yet, we did not find in our simulations a combination, which consistently provided improved performance. The incorporation of several eigenvectors for voice activity detection will be addressed in a future

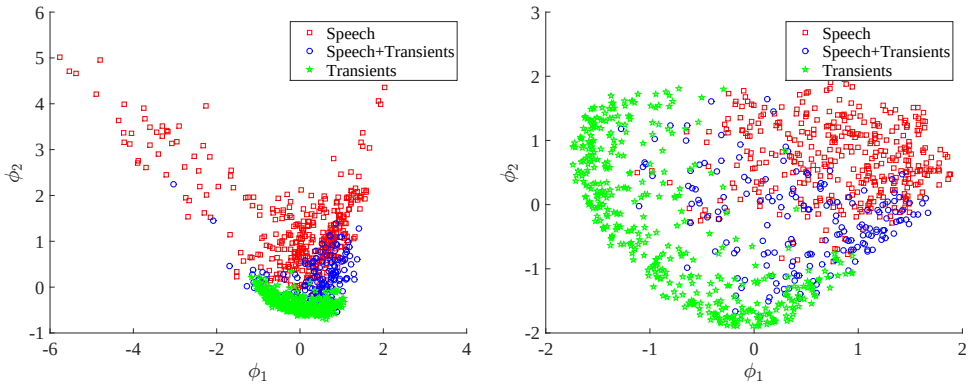


Figure 7.1: Scatter plot of the first two non-trivial eigenvectors, for which the speech signal is contaminated by a door-knocks transient. (left) Kernel based on the Euclidean distance and (right) kernel based on the modified Mahalanobis distance.

study.

An example of the obtained voice activity detection of a speech signal contaminated with keyboard taps transients is presented in Fig. 7.2 and Fig. 7.3. In Fig. 7.2, we qualitatively compare the performance of the proposed detector to the one presented in [99], which we term "Mousazadeh" in the plots. For both detectors, we set the threshold value to provide 90 percent correct detection rate and compare between their false alarms. Figure 7.2 (top) demonstrates that the false alarm rate of Mousadazeh is significantly higher than the false alarm rate of the proposed detector, especially in the non-speech region after the 15th second. We note that the method presented in [99] is based on representing the noisy signal using MFCCs, and then inferring a low-dimensional representation based on the Euclidean distance in which transient frames tend to be similar to speech frames as demonstrated in Fig. 7.1 (left). Therefore, it only partially distinguishes speech from transients, whereas the proposed method, based on the improved metric, provides

a better distinction between them.

In addition, the method presented in [99] is based on a supervised learning procedure in which the low-dimensional representation is obtained using a training set, and the transients are assumed known in advance. To make a fair comparison, in our simulations, we train the algorithm presented in [99] using several types of transients. In particular, for the evaluation of the algorithm, we use the same types of transients as in the training procedure, but the transients are taken from different recordings than those used for training. In contrast to Mousadazeh, the proposed method performs in an unsupervised manner, and the voice activity measure is learned from the sequence without any prior information.

To further gain insight into the voice activity detection obtained using the leading eigenvector $\boldsymbol{\phi}_1$ in (7.20), we plot the trajectory of $\boldsymbol{\phi}_1$ over time in Fig. 7.3. For the clarity of the presentation, we normalize the eigenvector in the plot to the range of 0 to 1. In addition, we recall that the eigenvector is used as a voice activity measure only for frames containing speech, transients or both of them; silent frames are assumed known in advance and they are assigned with the value zero in the plot. Figure 7.3 demonstrates that entries of the eigenvector with large values correspond to frames containing speech. Indeed, by setting the threshold to a value that yields 90 percent correct detection rate, the entries of the eigenvector, $\{\phi_1(n)\}$, that correspond to non-speech frames containing transients, receive values below the threshold. As a result, they correctly indicate absence of speech.

In addition to the method presented in [99], the performance of the proposed method is compared to the performance of the methods presented in [108, 123], and [62], which we term "Sohn", "Ramirez" and "Ishizuka" in the plots, respectively. The proposed method is termed "Proposed (MK)", where

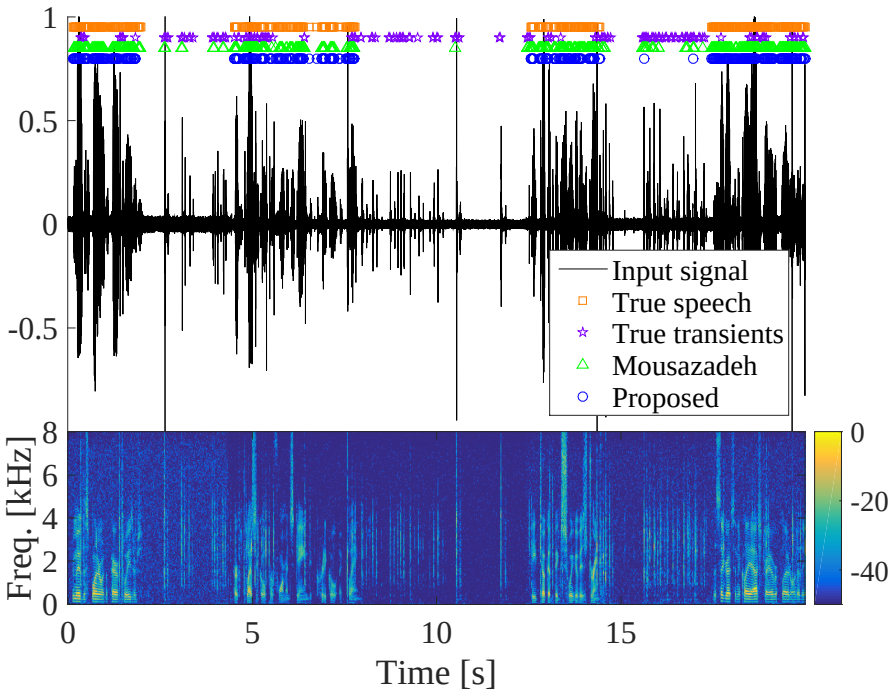


Figure 7.2: Qualitative assessment of the proposed VAD, with a keyboard taps transient. (Top) Time domain, input signal- black solid line, true speech- orange squares, true transients- purple stars, Mousazadeh with a threshold set for 90 percents correct detection rate- green triangles, proposed algorithm with a threshold set for 90 percent correct detection rate- blue circles. (Bottom) Spectrogram of the input signal.

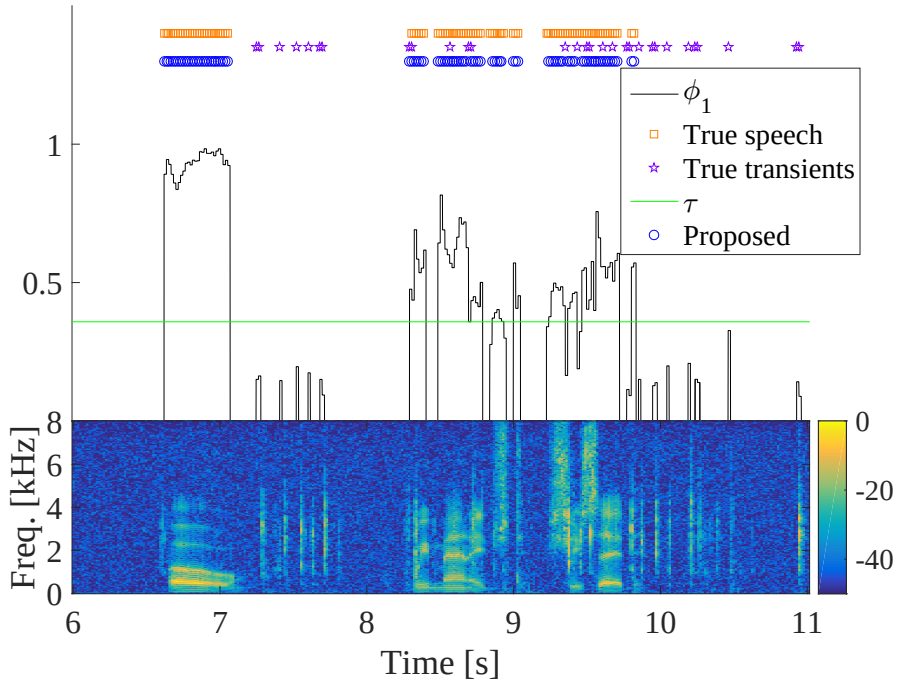


Figure 7.3: Qualitative assessment of the proposed VAD, with a keyboard taps transient. (Top) Time domain, the voice activity measure, i.e.,ϕ_1- black solid line, true speech- orange squares, true transients- purple stars, a threshold value τ providing 90 percents correct detection rate- green line, proposed algorithm with a threshold set for 90 percent correct detection rate- blue circles. (Bottom) Spectrogram of the input signal.

(MK) is the Mahalanobis kernel, and it is also compared to a similar method based on the Euclidean distance termed "Proposed (EK)", where (EK) is the Euclidean kernel. To better appreciate the results, we report on the delays induced by each method. The methods Sohn and Ishizuka operate in an online manner without a delay; Mousazadeh and Ramirez operate with a delay of two and four frames, respectively; and in the presented experiments, the proposed method operate in a batch manner. Yet, as already noted, the proposed method may be implemented in an online manner without a delay using a calibration set, given in advance without labels, similarly to the method we presented in [47].

The performance of the methods is evaluated in Figs. 7.4–7.6 and in Table 7.1. In Figs. 7.4–7.6 the methods are evaluated in the form of ROC curves, which are curves of probability of detection versus probability of false alarm. The ROC curves are generated by sweeping the threshold value in (7.21) from the minimal to the maximal entry of the leading eigenvector such that the higher the threshold is, the lower the correct detection and the false alarm rates are. The larger the AUC, the better the performance of the method; the AUC of each method is given in the legend box of each plot. Each of the Figs. 7.4–7.6 illustrates the performance of the methods for different types of transients: keyboard taps, hammering and door-knocks, respectively.

It can be seen in Figs. 7.4–7.6 that the competing methods Sohn, Ramirez and Ishizuka provide poor performance in distinguishing speech from transient frames since they are not designed for this particular task. In Figs. 7.4 and 7.5, the proposed method with the Euclidean distance and the method Mousadazeh, which both exploit the Euclidean distance to obtain a low-dimensional representation are comparable and perform significantly better than the methods presented in [108, 123] and [62]. Moreover, the proposed method based on the modified Mahalanobis distance provides the best performance. In Fig. 7.6, the proposed method provides comparable results to

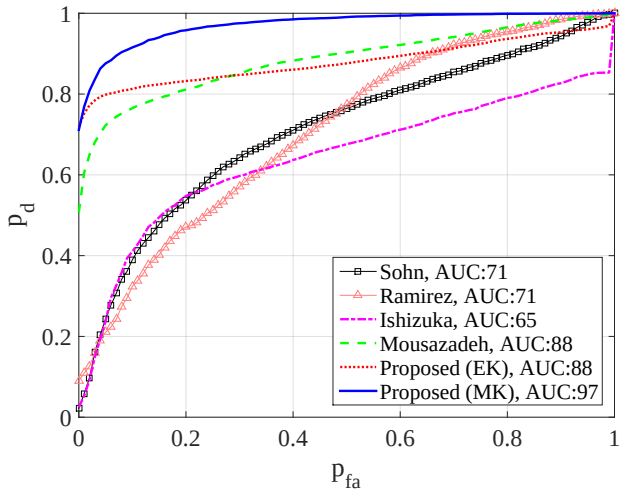


Figure 7.4: Probability of detection vs probability of false alarm. Test for a keyboard taps transient.

Mousadazeh and significantly outperforms all other methods.

We evaluate the proposed method for different ratios between the transients and speech, and report the AUC obtained for each method for different types of transients in Table 7.1. We define the transient to speech ratio as the ratio between the maximal amplitudes of the transients and speech such that for equal maximal amplitudes, as considered in the previous experiments, the ratio is one. To provide a fair comparison, the Mousadazeh method, which is the only method in our experiments based on supervised learning, is trained only for transient to speech ratio of 1 such that the transient to noise ratio is assumed to be unknown for all methods. We observe in Table 7.1 that for transient to speech ratio of 0.5, the proposed method

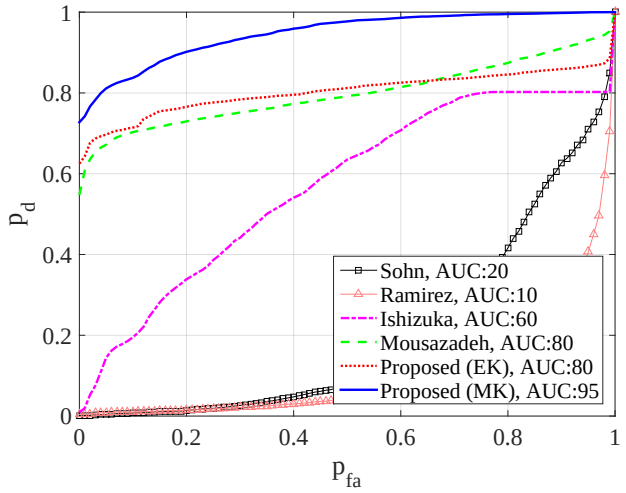


Figure 7.5: Probability of detection vs probability of false alarm. Test for a hammering transient.

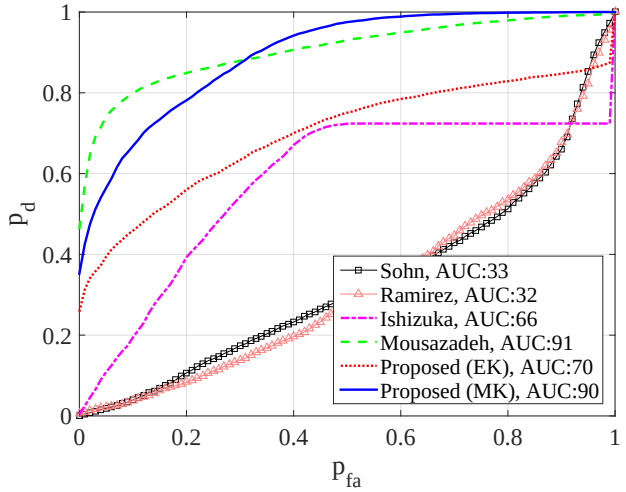


Figure 7.6: Probability of detection vs probability of false alarm. Test for a door-knocks transient.

based on the Mahalanobis distance provides performance, which is comparable to Mousazadeh and outperforms the competing detectors. Moreover, the proposed method provides the best performance in most of the experiments for transient to speech ratios of 1 and 2. The improved performance of the proposed method for high transient to speech ratio demonstrates its ability to reduce the dominance of the transients. In Table 7.1 (d) we summarize the average performance of the different methods for the different transients and transient to speech ratios demonstrating that the proposed method based on the modified Mahalanobis distance outperforms the competing detectors.

7.6 Conclusions

We have addressed the problem of voice activity detection in the presence of transients and have proposed a modified version of the Mahalanobis distance, which better distinguishes between speech and transients. To motivate the use of the modified Mahalanobis distance, we have presented a model in which speech and transients are represented by two independent sets of generating variables. The generating variables represent the content of the signal, i.e., speech and transients, and, as a result, speech and transients are successfully distinguished. Although the generating variables are not directly accessible, we have shown that distances between them can be approximated by the modified Mahalanobis distance. Moreover, we have shown that the Mahalanobis distance approximates the Euclidean distance between re-scaled variables for which the dominance of the transients is reduced; therefore, it is especially suitable for voice activity detection since it allows us to cluster the signal according to the presence of speech rather than according to the presence of transients. The main limitation in the use of the Mahalanobis distance is that the approximation of the Euclidean distance between the generating variables holds only for small distances. To overcome this prob-

	Keyboard taps	Hammering	Door-knocks	Metronome	Scissors	Crackles
Sohn	71	20	33	28	68	59
Ramirez	71	10	32	14	65	60
Ishizuka	65	60	66	57	74	49
Mousazadeh	88	80	**91**	80	87	**93**
Proposed (EK)	88	80	70	81	**93**	79
Proposed (MK)	**97**	**95**	90	**91**	**93**	88

(a)

	Keyboard taps	Hammering	Door-knocks	Metronome	Scissors	Crackles
Sohn	52	14	26	16	48	39
Ramirez	49	6	25	6	42	38
Ishizuka	63	53	64	57	73	51
Mousazadeh	71	58	81	58	64	76
Proposed (EK)	91	79	72	82	**96**	78
Proposed (MK)	**97**	**91**	**87**	**91**	92	**86**

(b)

	Keyboard taps	Hammering	Door-knocks	Metronome	Scissors	Crackles
Sohn	86	29	43	48	83	79
Ramirez	87	20	45	35	85	81
Ishizuka	66	63	66	58	74	45
Mousazadeh	94	86	**95**	**90**	**92**	**96**
Proposed (EK)	86	82	77	81	90	80
Proposed (MK)	**95**	**97**	92	86	91	90

(c)

Sohn	Ramirez	Ishizuka	Mousazadeh	Proposed (EK)	Proposed (MK)
47	43	61	82	83	**92**

(d)

Table 7.1: (a) AUC scores; transient to speech ratio: 1. (b) AUC scores; transient to speech ratio: 2. (c) AUC scores; transient to speech ratio: 0.5. (d) Average AUC scores.

lem, we have proposed to exploit a kernel-based manifold learning approach that integrates short Mahalanobis distances into a global canonical representation of the signal. We have shown that the canonical representation successfully divides the signal into speech and non-speech clusters. Based on the canonical representation we have proposed a measure of voice activity providing improved performance compared to competing detectors.

Chapter 8

Research Summary And Future Research Directions

8.1 Research Summary

In this thesis, we addressed the processing of multi-modal signals comprising multiple sources of data. We considered a setting, in which part of the sources are common to the different modalities and are considered as sources of interest such as speech of a particular speaker, while other sources such as non-speech signals are considered interferences. In this setting, we addressed the major open questions of data fusion, the processing of data which is only partially available across the modalities and of multi-modal correspondence. We approached these questions via manifold learning by analyzing the combination of single-modal kernels. Both via graph-theoretic analysis and via various audio-visual applications, we have shown that the local relations (affinities) between data points, as are defined by the single-modal and the combined kernels, play a critical roll in these open questions. In particular, we have shown how to address these questions via the unique combination of the signals by a product of kernels, which provides a useful representation according to the common source reducing the effect of the interfering sources.

Based on this approach we explored challenging applications related to the analysis of audio-visual sound scenes including sound source activity detection and audio localization in video. Experimental results demonstrated that the proposed kernel-based geometric methods provide improved performance for the various tasks. The main contributions of the thesis chapters are as follows:

In Chapter 3, we addressed the problem of sensor fusion and proposed to fuse the multi-modal signals by the product of affinity kernels constructed separately for each modality. We introduced a graph-theoretic analysis that quantifies the relations between single and multi-modal affinities by linking between them and the connectivity of the corresponding single and multi-modal graphs. We showed that the fusion process increases the connectivity of the multi-modal graph. Accordingly, selecting the single-modal kernels as in the single modal case, which, to the best of our knowledge, is the practice in all previous studies, leads to a redundant connectivity in the multi-modal graph. Based on the proposed model, we proposed to reduce this connectivity by the proper selection of the single-modal kernel bandwidths. We presented an algorithm for their selection, by relating the connectivity of the multi-modal kernel to the single modal kernels as if the latter are constructed in a single modal setting.

We demonstrated the improved performance of the proposed fusion approach for audio-visual voice activity detection in the presence of various challenging interfering sources such as high levels of background noise, transient interferences and even speech from other speakers. The latter, which was addressed in Chapter 4, is considered a particularly challenging source since both the desired and the interfering sources are speech and they share similar characteristics. We showed that the proposed fusion approach allows for a good representation of the common source while reducing the effect of the interfering sources and providing improved performance compared to competing methods.

We further showed in Chapter 3 that reducing the connectivity of the multi-modal graph, by a proper selection of the multi-modal kernel bandwidth further improves the fusion process and the algorithm for their selection leads to near optimal values. These findings implicitly imply that to obtain a representation of the data according to the geometry of the common source, the connectivity of the multi-modal graph should be kept as small as possible while maintaining the necessary condition of a connected graph. This is probably because samples containing the common source often appear similar (affine) in content even locally to samples containing the other sources.

In Chapter 5, we extended our setting to an online case where data from the different modalities is available only in certain time intervals. We considered this setting in the context of sound scene analysis, in which one would like to detect the activity of more than one source, but since only a single video camera is assumed, video recordings of the sound sources are available for only short time intervals. We showed that the geometry of the common source can be successfully learned from a short time interval and the obtained representation can be extended to new time intervals, even when only one modality is available. Experimental results showed that the proposed approach outperforms the single modal approach based only on the audio signal demonstrating the usefulness of incorporating multi-modal signals even if they only partially available. Moreover, our experimental results demonstrated the challenge in incorporating the video signal, which contains high levels of interfering sources such that using merely the video signal or incorporating it with the audio using competing fusion approaches led to poor activity detection performance compared to the proposed approach.

In Chapter 6, we explored the question of how to measure correspondence between multi-modal signals. We proposed to use the trace of the kernel product as a measure of multi-modal correspondence and motivated its use by revisiting the graph-theoretic model presented in Chapter 3. We showed

that higher values of the measure are expected for signals in the different modalities, which correspond to each other since the connectivity of their corresponding graphs is expected to be correlated. We then addressed the online setting and proposed a method for the efficient update of the measure for new incoming frames. Not only the proposed measure does not require eigendecomposition, but we further showed that it can be computed with the computation complexity of $O(N)$, while completely avoiding matrix multiplication, whose complexity is higher than $O\left(N^2\right)$. Experimental results have shown for the applications of audio localization in video and eye-fixation prediction that the proposed measure properly reduces modality-specific factors and provides improved performance compared to competing methods.

Last, Chapter 7 deals with the question of how to properly distinguish between different sources of the signal. In the single modal setting, we considered the task of voice activity detection in the presence of transient interferences. The main challenge in this task is that the different sources, i.e., speech and transients, wrongly appear similar through, e.g., the Euclidean distance. Similarly to the multi-modal case, the key to successful separation between the sources via manifold learning is the design of an affinity kernel which properly measures affinities between samples of the signals. Accordingly, we proposed an affinity kernel, which is based on a modified Mahalanobis distance exploiting the difference in the variation rate of the sources over time. To motivate the use of this kernel, we proposed an analysis based on modeling the speech and the transients by hidden variables controlling their generation. We showed that the proposed kernel attenuates the transients which are assumed to be fast varying and demonstrated improved voice activity detection compared to competing methods in the presence of various types of transients.

8.2 Future Research Directions

Throughout this thesis, we showed various cases where multi-modal signals comprise multiple sources of data, which can be considered modality-specific or common across the modalities. This observation raises interesting questions that are related to the interplay between the sources. In the context of the kernel product, the analysis in the literature based on diffusion geometry suggests that the modality-specific sources are completely reduced by the fusion process. However, it can be observed in our experiments that the ratio between the common and the modality-specific sources, i.e., between the desired signal and the interferences, has an important role on the representation of the data. Accordingly, quantifying this ratio and understanding its effect on the fusion process are open questions. The proposed graph-theoretic analysis of both the fusion process and the correspondence measure may be a convenient starting point for exploring these questions since it provides insights about the influence of the intensities of the modality-specific interferences. For example, in the fusion process, we expect that high levels of interferences will lead to wrong connections between samples. Similarly, we expect low correspondence levels between modalities of signals with high levels of interferences.

A related question is how to obtain a representation of the signal according to the modality-specific sources, which is complementary to the representation of the common source. These two representations may be jointly used, for example, for the application of signal enhancement in the presence of multiple sound sources considered in Chapter 5. Indeed, speech enhancement algorithms typically exploit activity detection of speech and the interferences for the estimation of their corresponding spectrums, which are, in turn, used for filtering. A possible approach would be to design a complementary kernel to the kernel of products, which is used for the representation of the common source. While the graph-theoretic analysis implies that the product of kernels increases the connectivity of the graph with respect to the single modal

kernel, the modality-specific kernel should reduce this connectivity.

Another open question is how to fuse data measured by more than two modalities. While one can combine multiple kernels, e.g., via the product between them, the common source to *all* modalities might not be useful. For example in the fusion of electroencephalography (EEG) signals, some sensors may not contain the desired source of data such that fusing them along with other sources may provide a representation where the desired source is suppressed [66]. Accordingly, a related question would be how to choose a set of sensors that are most suitable for the fusion process. This question is related back to the question of quantifying the ratio between the common source and the interferences such that only modalities that have low levels of modality-specific sources will be taken into account.

Finally, these questions should be considered in an online setting. Beyond the technical consideration of how to obtain a representation given a new sample, it is important to consider the variation of the geometry of the data over time [116]. For example, the geometry learned from a sequence of samples may not be relevant after a certain time if new sources are present.

Bibliography

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

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

[3] *[Online]. Available: http://www.freesound.org.*

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

[5] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe, "Analyzing free-standing conversational groups: A multimodal approach," in *Proc. the 23rd ACM International Conference on Multimedia*, ser. MM '15. New York, NY, USA: ACM, 2015, pp. 5–14. [Online]. Available: http://doi.acm.org/10.1145/2733373.2806238

[6] I. Almajai and B. Milner, "Using audio-visual features for robust voice activity detection in clean and noisy speech," in *Proc. 16th European Signal Processing Conference*. IEEE, 2008, pp. 1–5.

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

[8] P. Atrey, M. Hossain, A. El Saddik, and M. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[9] A. Aubrey, B. Rivet, Y. Hicks, L. Girin, J. Chambers, and C. Jutten, "Two novel visual voice activity detectors based on appearance models and retinal filltering," in *Proc. 15th European Signal Processing Conference (EUSIPCO), 2007*, 2007.

[10] A. Aubrey, Y. Hicks, and J. Chambers, "Visual voice activity detection with optical flow," *IET Image Processing*, vol. 4, no. 6, pp. 463–472, 2010.

[11] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," 2005.

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

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

[14] J. P. Barker and F. Berthommier, "Evidence of correlation between acoustic and visual features of speech," in *Proc. Int. Congress of Phonetical Sciences*, 1999, pp. 199–202.

[15] J. Barron, D. Fleet, and S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, no. 1, pp. 43–77, 1994.

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

[17] 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.

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

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

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

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

[22] 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.

[23] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, July 2015, pp. 1–7.

[24] A. L. Casanovas and P. Vandergheynst, "Audio-based nonlinear video diffusion," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2010, pp. 2486–2489.

[25] J. H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 1965–1976, 2006.

[26] J. H. Chang, J. W. Shin, and N. S. Kim, "Likelihood ratio test with complex laplacian model for voice activity detection." in *Proc. the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2003.

[27] J. Chang and N. Kim, "Voice activity detection based on complex laplacian model," *Electronics Letters*, vol. 39, no. 7, pp. 632–634, 2003.

[28] S. Chu, S. Narayanan, and C. J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[29] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[30] I. Cohen and B. Berdugo, "Spectral enhancement by tracking speech presence probability in subbands," in *Proc. IEEE Workshop on Hands-Free Speech Communication, HSC'01, 2001*, pp. 95–98.

[31] ——, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[32] ——, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, 2002.

[33] R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer, "Graph laplacian tomography from unknown random projections," *IEEE Transactions on Image Processing*, vol. 17, no. 10, pp. 1891–1899, 2008.

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

[35] 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.

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

[37] M. Cristani, M. Bicego, and V. Murino, "Audio-visual event recognition in surveillance video sequences," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 257–267, 2007.

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

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

[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] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[42] 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.

[43] J. Dennis, H. Tran, and E. S. Chng, "Image feature representation of the subband power distribution for robust sound event classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 367–377, 2013.

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

[45] 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.

[46] D. Dov and I. Cohen, "Voice activity detection in presence of transients using the scattering transform," in *Proc. IEEE 28th Convention of Electrical & Electronics Engineers in Israel (IEEEI)*, 2014, pp. 1–5.

[47] D. Dov, R. Talmon, and I. Cohen, "Audio-visual voice activity detection using diffusion maps," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 732–745, 2015.

[48] ——, "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.

[49] ——, "Kernel method for speech source activity detection in multimodal signals," in *Proc. IEEE International Conference on the Science of Electrical Engineering (ICSEE),*. IEEE, 2016, pp. 1–5.

[50] ——, "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.

[51] ——, "Sequential audio-visual correspondence with alternating diffusion kernels," *IEEE Transactions on Signal Processing*, vol. 66, no. 12, pp. 3100–3111, 2018.

[52] ——, "Kernel method for voice activity detection in the presence of transients," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2313–2326, 2016.

[53] C. J. Dsilva, R. Talmon, C. W. Gear, R. R. Coifman, and I. G. Kevrekidis, "Data-driven reduction for multiscale stochastic dynamical systems," *arXiv preprint arXiv:1501.05195*, 2015.

[54] 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. Advances in Neural Information Processing Systems (NIPS)*, 2000, pp. 772–778.

[55] 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.

[56] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic-phonetic continous speech database," National Inst. of Standards and Technology (NIST), Gaithersburg, MD, Feb 1993.

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

[58] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[59] A. Hirszhorn, D. Dov, R. Talmon, and I. Cohen, "Transient interference suppression in speech signals based on the OM-LSA algo-

rithm," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

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

[62] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, "Noise robust voice activity detection based on periodic to aperiodic component ratio," *Speech Communication*, vol. 52, no. 1, pp. 41–60, 2010.

[63] G. Iyengar, H. J. Nock, and C. Neti, "Audio-visual synchrony for detection of monologues in video archives," in *Proc. International Conference on Multimedia and Expo (ICME)*, vol. 1. IEEE, 2003, pp. I–329.

[64] 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.

[65] I. Jhuo, G. Ye, S. Gao, D. Liu, Y. Jiang, D. Lee, and S. Chang, "Discovering joint audio–visual codewords for video event detection," *Machine Vision and Applications*, vol. 25, no. 1, pp. 33–47, 2014.

[66] O. Katz, R. Talmon, Y. Lo, and H. Wu, "Diffusion-based nonlinear filtering for multimodal data fusion with application to sleep stage assessment," *arXiv preprint arXiv:1701.03619*, 2017.

[67] 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.

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

[69] ——, "Cross-modal localization via sparsity," *IEEE Transactions on Signal Processing*, vol. 55, no. 4, pp. 1390–1404, 2007.

[70] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H. Li, "Voice activity detection using mfcc features and support vector machine," in *Proc. Int. Conf. on Speech and Computer (SPECOM07), Moscow, Russia*, vol. 2, 2007, pp. 556–561.

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

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

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

[74] D. Kushnir, "Active-transductive learning with label-adapted kernels," in *Proc. the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '14. New York, NY, USA: ACM, 2014, pp. 462–471. [Online]. Available: http://doi.acm.org/10.1145/2623330.2623673

[75] D. Kushnir, A. Haddad, and R. R. Coifman, "Anisotropic diffusion on sub-manifolds with application to earth structure classification," *Applied and Computational Harmonic Analysis*, vol. 32, no. 2, pp. 280–294, 2012.

[76] 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 Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1854–1864, Oct 2016.

[77] S. Lafon, Y. Keller, and R. Coifman, "Data fusion and multicue data matching by diffusion maps," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1784–1797, 2006.

[78] 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.

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

[80] 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.

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

[82] O. Lindenbaum, A. Yeredor, and M. Salhov, "Learning coupled embedding using multiview diffusion maps," in *Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 127–134.

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

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

[85] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. 1st International Conference on Music Information Retrieval (ISMIR)*, 2000.

[86] P. C. Mahalanobis, "On the generalized distance in statistics," *Proceedings of the National Institute of Sciences (Calcutta)*, vol. 2, pp. 49–55, 1936.

[87] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," 1976.

[88] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016. [Online]. Available: http://www.mdpi.com/2076-3417/6/6/162

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

[90] T. Michaeli, W. Wang, and T. Livescu, "Nonparametric canonical correlation analysis," *Submitted to International Conference on Learning Representations (ICLR 2016)*.

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

[92] 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, p. 6, 2016.

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

[94] V. Minotto, C. Lopes, J. Scharcanski, C. Jung, and B. Lee, "Audio-visual voice activity detection based on microphone arrays and color information," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 147–156, 2013.

[95] X. A. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[96] 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.

[97] G. Mishne, R. Talmon, and I. Cohen, "Graph-based supervised automatic target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2738–2754, 2015.

[98] 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.

[99] S. Mousazadeh and I. Cohen, "Voice activity detection in presence of transient noise using spectral clustering." *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, no. 6, pp. 1261–1271, 2013.

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

[101] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 6440–6444.

[102] 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.

[103] 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.

[104] 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.

[105] L. Rabiner and B. Juang, "Fundamentals of speech recognition," 1993.

[106] J. Ramirez, J. M. Górriz, and J. C. Segura, "Voice activity detection. fundamentals and speech secognition system robustness," 2007.

[107] J. Ramírez, J. C. Segura, and J. M. Górriz, "Revised contextual LRT for voice activity detection," in *Proc. 32th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 2007, pp. IV–801.

[108] J. Ramìrez, J. Segura, C. Benìtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3, pp. 271–287, 2004.

[109] O. Rosen, S. Mousazadeh, and I. Cohen, "Voice activity detection in presence of transient noise using spectral clustering and diffusion kernels," in *Proc. IEEE 28th Convention of Electrical & Electronics Engineers in Israel (IEEEI), 2014.* IEEE, 2014, pp. 1–5.

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

[111] N. Seichepine, S. Essid, C. FÃ©votte, and O. CappÃ©, "Soft nonnegative matrix co-factorizationwith application to multimodal speaker diarization," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 3537–3541.

[112] ——, "Soft nonnegative matrix co-factorization," *IEEE Transactions on Signal Processing*, vol. 62, no. 22, pp. 5940–5949, Nov 2014.

[113] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[114] J. Shin, J. Chang, H. Yun, and N. Kim, "Voice activity detection based on generalized gamma distribution," in *Proc. the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, pp. 781–784.

[115] J. Shin, H. Kwon, S. Jin, and N. Kim, "Voice activity detection based on conditional map criterion," *IEEE Signal Processing Letters*, vol. 15, pp. 257–260, 2008.

[116] T. Shnitzer, R. Talmon, and J. J. Slotine, "Diffusion maps kalman filter," *arXiv preprint arXiv:1711.09598*, 2017.

[117] S. Siatras, N. Nikolaidis, M. Krinidis, and I. Pitas, "Visual lip activity detection and speaker detection using mouth region intensities," *IEEE*

Transactions on Circuits and Systems for Video Technology, vol. 19, no. 1, pp. 133–137, 2009.

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

[119] A. Singer and R. Coifman, "Non-linear independent component analysis with diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 25, no. 2, pp. 226–239, 2008.

[120] A. Singer, R. Erban, I. Kevrekidis, and R. Coifman, "Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps," *Proceedings of the National Academy of Sciences*, vol. 106, no. 38, pp. 16 090–16 095, 2009.

[121] D. Sodoyer, B. Rivet, L. Girin, C. Savariaux, J. Schwartz, and C. Jutten, "A study of lip movements during spontaneous dialog and its application to voice activity detection," *The Journal of the Acoustical Society of America*, vol. 125, p. 1184, 2009.

[122] D. Sodoyer, B. Rivet, L. Girin, J. Schwartz, and C. Jutten, "An analysis of visual speech information applied to voice activity detection," in *Proc. 31st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2006, pp. I–I.

[123] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[124] P. Somervuo, A. Härmä, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2252–2263, 2006.

[125] 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.

[126] S. Steinerberger, "A filtering technique for markov chains with applications to spectral embedding," *arXiv preprint arXiv:1411.1638*, 2014.

[127] B. H. Story, "A parametric model of the vocal tract area function for vowel and consonant simulation," *The Journal of the Acoustical Society of America*, vol. 117, no. 5, pp. 3231–3254, 2005.

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

[129] R. Talmon, I. Cohen, and S. Gannot, "Clustering and suppression of transient noise in speech signals using diffusion maps," in *Proc. 36th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5084–5087.

[130] ——, "Transient noise reduction using nonlocal diffusion filters," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1584–1599, 2011.

[131] ——, "Single-channel transient interference suppression with diffusion maps," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 132–144, 2013.

[132] R. Talmon, I. Cohen, S. Gannot, and R. Coifman, "Diffusion maps for signal processing: A deeper look at manifold-learning techniques based on kernels and graphs," *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 75–86, 2013.

[133] R. Talmon, I. Cohen, S. Gannot, and R. R. Coifman, "Supervised graph-based processing for sequential transient interference suppression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2528–2538, 2012.

[134] R. Talmon and R. R. Coifman, "Empirical intrinsic geometry for nonlinear modeling and time series filtering," *Proceedings of the National Academy of Sciences*, vol. 110, no. 31, pp. 12 535–12 540, 2013.

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

[136] S. Tamura, M. Ishikawa, T. Hashiba, S. T., and S. Hayamizu, "A robust audio-visual speech recognition using audio-visual voice activity detection," in *Proc. the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2010, pp. 2694–2697.

[137] P. Tiawongsombat, M. Jeong, J. Yun, B. You, and S. Oh, "Robust visual speakingness detection using bi-level HMM," *Pattern Recognition*, vol. 45, no. 2, pp. 783–793, 2012.

[138] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[139] H. D. Tran and H. Li, "Sound event recognition with probabilistic distance svms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1556–1568, 2011.

[140] X. Valero and F. Alias, "Gammatone cepstral coefficients: biologically inspired features for non-speech audio classification," *IEEE Transactions on Multimedia*, vol. 14, no. 6, pp. 1684–1689, 2012.

[141] 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.

[142] I. Volfin and I. Cohen, "Dominant speaker identification for multipoint videoconferencing," *Computer Speech & Language*, vol. 27, no. 4, pp. 895–910, 2013.

[143] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[144] 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, p. 1583, 2000.

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

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

[147] 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.

[148] C. Xu, C. Xiong, and J. Corso, "Streaming hierarchical video segmentation," *Proc. European Conference on Computer Vision (ECCV)*, pp. 626–639, 2012.

[149] 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.

[150] T. Yoshida, K. Nakadai, and H. G. Okuno, "An improvement in audio-visual voice activity detection for automatic speech recognition," in *Proc. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2010, pp. 51–61.

[151] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering." in *Proc. Advances in Neural Information Processing Systems (NIPS)*, vol. 17, no. 1601-1608, 2004, p. 16.

[152] 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. [Online]. Available: https://doi.org/10.1007/s11042-016-3294-5

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

Appendices

Appendix I

Second Order Taylor Expansion at The Mid-point

Recall that the mid-point $\mathbf{y}_p$ is given by $\mathbf{y}_p = \frac{\mathbf{y}_n + \mathbf{y}_m}{2}$; by a second order Taylor expansion at the mid-point, the ith re-scaled generating variable, denoted by $\boldsymbol{\psi}_n(i)$, is given by [75]:

$$\begin{aligned}\boldsymbol{\psi}_n(i) = \boldsymbol{\psi}_p(i) + \tfrac{1}{2}\textstyle\sum_j g_{i,j}\left(\mathbf{y}_p\right)\left(\mathbf{y}_n(j) - \mathbf{y}_m(j)\right) \\ + \tfrac{1}{8}\textstyle\sum_{kl} g_{i,kl}\left(\mathbf{y}_p\right)\left(\mathbf{y}_n(k) - \mathbf{y}_m(k)\right)\left(\mathbf{y}_n(l) - \mathbf{y}_m(l)\right) + O\left(\left\|\mathbf{y}_n - \mathbf{y}_m\right\|^3\right).\end{aligned} \tag{8.1}$$

where g_i is the ith element in g, $g_{i,j} \triangleq \frac{\partial g_i}{\partial \mathbf{y}_n(j)}$ and $g_{i,kl} \triangleq \frac{\partial g_i}{\partial \mathbf{y}_n(k)\partial \mathbf{y}_n(l)}$. Using a similar expansion of $\boldsymbol{\psi}_m(i)$ around the mid-point, the Euclidean distance between the re-scaled variables is given by:

$$\begin{aligned}\left\|\boldsymbol{\psi}_n - \boldsymbol{\psi}_m\right\|^2 = \textstyle\sum_i \left(\boldsymbol{\psi}_n(i) - \boldsymbol{\psi}_m(i)\right)^2 = \\ \textstyle\sum_{ijk} g_{i,j}\left(\mathbf{y}_p\right) g_{i,k}\left(\mathbf{y}_p\right)\left(\mathbf{y}_n(j) - \mathbf{y}_m(j)\right)\left(\mathbf{y}_n(k) - \mathbf{y}_m(k)\right) + O\left(\left\|\mathbf{y}_n - \mathbf{y}_m\right\|^4\right),\end{aligned}$$

because all multiplications terms comprising the second order of the Taylor expansion in (8.1) are of the order of $O\left(\left\|\mathbf{y}_n - \mathbf{y}_m\right\|^4\right)$ due to symmetry. In a matrix notation, we have:

$$\|\boldsymbol{\psi}_n - \boldsymbol{\psi}_m\|^2 = (\mathbf{y}_n - \mathbf{y}_m)^T \boldsymbol{\Lambda}^{-1}(\mathbf{y}_p)(\mathbf{y}_n - \mathbf{y}_m) + O\left(\|\mathbf{y}_n - \mathbf{y}_m\|^4\right),$$

where $\boldsymbol{\Lambda}\left(\boldsymbol{\psi}_p\right) \triangleq \mathbf{J}\mathbf{J}^T\left(\boldsymbol{\psi}_p\right) \in \mathbb{R}^{L\times L}$, and $\mathbf{J}\left(\boldsymbol{\psi}_p\right) \in \mathbb{R}^{L\times d}$ is the Jacobian of the function h at the mid-point.

Appendix II

Jacobian at The Mid-Point

Let $\gamma_{ij}(\mathbf{y}_n)$ be the (i,j)th entry of $\boldsymbol{\Lambda}^{-1}(\mathbf{y}_n)$; the first order Taylor expansions of $\gamma_{ij}(\mathbf{y}_n)$ and $\gamma_{ij}(\mathbf{y}_m)$ at the mid-point are given by:

$$\gamma_{ij}(\mathbf{y}_n) = \gamma_{ij}(\mathbf{y}_p) + \frac{1}{2}\sum_k \gamma_{ij,k}(\mathbf{y}_p)(\mathbf{y}_n(k) - \mathbf{y}_m(k)) + O\left(\|\mathbf{y}_n - \mathbf{y}_m\|^2\right),$$

$$\gamma_{ij}(\mathbf{y}_m) = \gamma_{ij}(\mathbf{y}_p) + \frac{1}{2}\sum_k \gamma_{ij,k}(\mathbf{y}_p)(\mathbf{y}_m(k) - \mathbf{y}_n(k)) + O\left(\|\mathbf{y}_m - \mathbf{y}_n\|^2\right),$$

where $\gamma_{ij,k}(\mathbf{y}_n) = \frac{\partial \gamma_{ij}}{\partial \mathbf{y}_n(k)}$. The summation of this two equations yields:

$$\gamma_{ij}(\mathbf{y}_p) = \frac{1}{2}\gamma_{ij}(\mathbf{y}_n) + \frac{1}{2}\gamma_{ij}(\mathbf{y}_m) + O\left(\|\mathbf{y}_n - \mathbf{y}_m\|^2\right).$$

Hence, in a matrix form, we have:

$$\boldsymbol{\Lambda}^{-1}(\mathbf{y}_p) = \frac{1}{2}\boldsymbol{\Lambda}^{-1}(\mathbf{y}_n) + \frac{1}{2}\boldsymbol{\Lambda}^{-1}(\mathbf{y}_m) + O\left(\|\mathbf{y}_n - \mathbf{y}_m\|^2\right),$$

and by substituting the last equation into (7.13), we have:

$$\|\boldsymbol{\psi}_n - \boldsymbol{\psi}_m\|^2 = \tfrac{1}{2}\left(\mathbf{y}_n - \mathbf{y}_m\right)^T \left(\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_n\right) + \boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_m\right)\right)\left(\mathbf{y}_n - \mathbf{y}_m\right) + O\left(\|\mathbf{y}_n - \mathbf{y}_m\|^4\right).$$

Appendix III

Local Jacobian vs Local Covariance

We estimate the covariance matrix $\mathbf{C}_n$ of the observable signal $\mathbf{y}_n$ at a small temporal neighborhood of frame n, and assume a set of frames in a small temporal neighborhood of $\mathbf{y}_n$ for which $\mathbf{y}_n$ is the mean value. The first order Taylor expansion of an arbitrary frame, denoted by $\mathbf{y}$, around $\mathbf{y}_n$ is given by:

$$\mathbf{y} = \mathbf{y}_n + \mathbf{J}\left(\boldsymbol{\psi}_n\right)\left(\boldsymbol{\psi} - \boldsymbol{\psi}_n\right) + O\left(\|\boldsymbol{\psi} - \boldsymbol{\psi}_n\|^2\right). \tag{8.2}$$

The relation between the covariance matrix $\mathbf{C}_n$ and the Jacobian $\mathbf{J}\left(\boldsymbol{\psi}_n\right)$ in frame n is given by:

$$\begin{aligned}\mathbf{C}_n &= \mathbb{E}\left[\left(\mathbf{y} - \mathbf{y}_n\right)\left(\mathbf{y} - \mathbf{y}_n\right)^T\right] \\ &= \mathbf{J}\left(\boldsymbol{\psi}_n\right)\mathbb{E}\left[\left(\boldsymbol{\psi} - \boldsymbol{\psi}_n\right)\left(\boldsymbol{\psi} - \boldsymbol{\psi}_n\right)^T\right]\mathbf{J}^T\left(\boldsymbol{\psi}_n\right) + O\left(\|\boldsymbol{\psi} - \boldsymbol{\psi}_n\|^3\right) \\ &= \boldsymbol{\Lambda}\left(\boldsymbol{\psi}_n\right) + O\left(\|\boldsymbol{\psi} - \boldsymbol{\psi}_n\|^3\right),\end{aligned} \tag{8.3}$$

where assuming that $\boldsymbol{\psi}_n$ is the mean value of the generating variables, $\mathbb{E}\left[\left(\boldsymbol{\psi} - \boldsymbol{\psi}_n\right)\left(\boldsymbol{\psi} - \boldsymbol{\psi}_n\right)^T\right]$ is the covariance of $\boldsymbol{\psi}$, which is the identity matrix due to the normalization in (7.8) and (7.9). We note that the error term in (8.3) is neglected since we assume that frames in a small temporal neighborhood tend to be more similar to $\mathbf{y}_n$ compared to an arbitrary frame $\mathbf{y}_m$. Moreover, assuming a symmetric distribution of $\boldsymbol{\psi}$ around $\boldsymbol{\psi}_n$, e.g. a Gaussian distribution, the error term becomes of the order of four since odd moments of the distribution equal zero. By following the derivation pre-

sented in [75], it may be further shown that $\left[\mathbf{\Lambda}\left(\boldsymbol{\psi}_n\right)+O\left(\|\boldsymbol{\psi}-\boldsymbol{\psi}_n\|^3\right)\right]^{-1}=\left[\mathbf{\Lambda}^{-1}\left(\mathbf{y}_n\right)+O\left(\|\mathbf{y}-\mathbf{y}_n\|^3\right)\right]$. This result is obtained by further assuming that the function f in (7.4) is bi-Lipschitz such that distances between frames in the observable domain $\|\mathbf{y}-\mathbf{y}_n\|$ are of the same order as in the domain of the generating variables $\|\boldsymbol{\psi}-\boldsymbol{\psi}_n\|$. Hence, by setting $\mathbf{C}_n^{-1}\approx\mathbf{\Lambda}^{-1}\left(\mathbf{y}_n\right)$ in (7.13) we have:

$$\|\mathbf{y}_n-\mathbf{y}_m\|_M^2=\|\boldsymbol{\psi}_n-\boldsymbol{\psi}_m\|^2+O\left(\|\mathbf{y}_n-\mathbf{y}_m\|^4\right).$$

Appendix IV

Modified Mahalanobis Distance for Different Hypotheses

The derivation of (7.6) was made in Section 7.3 under the assumption that hypotheses $\mathcal{H}_1^x$ and $\mathcal{H}_1^t$ hold for both frames $\mathbf{y}_n$ and $\mathbf{y}_m$. We now address the other hypotheses, starting from the case in which only speech is present in both $\mathbf{y}_n$ and $\mathbf{y}_m$, and as a result, in the mid-point $\mathbf{y}_p$ as well. Since both frames are independent of transients, the partial derivatives of entries of the function g with respect to the generating variables of transients equal zero, i.e., $\forall j: g_{i,j}=\frac{\partial g_i}{\partial y_n(j)}=0$. Accordingly, the Jacobian of g is reduced to:

$$\mathbf{J}\left(\boldsymbol{\psi}_n\right)\triangleq\left[\begin{array}{c}\mathbf{J}_x\left(\boldsymbol{\psi}_n^x\right)\\ \mathbf{J}_t\left(\boldsymbol{\psi}_n^t\right)\end{array}\right]=\left[\begin{array}{c}\mathbf{J}_x\left(\boldsymbol{\psi}_n^x\right)\\ 0\end{array}\right],$$

where $\mathbf{J}_x\left(\boldsymbol{\psi}_n^x\right)\in\mathbb{R}^{d^x\times L}$ and $\mathbf{J}_t\left(\boldsymbol{\psi}_n^t\right)\in\mathbb{R}^{d^t\times L}$ are the parts of the Jacobian associated with the generating variables of the speech and of the transients, respectively, and hence:

$$\mathbf{\Lambda}\left(\boldsymbol{\psi}_n\right)\triangleq\mathbf{J}\mathbf{J}^{\mathbf{T}}\left(\boldsymbol{\psi}_n\right)=\mathbf{J}_x\mathbf{J}_x^T\left(\boldsymbol{\psi}_n^x\right). \tag{8.4}$$

By substituting (8.4) into (7.13) and using a similar derivation as in Appendix II, we obtain a result similar to (7.14):

$$\|\boldsymbol{\psi}_n - \boldsymbol{\psi}_m\|^2 = \tfrac{1}{2}\left(\mathbf{y}_n - \mathbf{y}_m\right)^T \left(\boldsymbol{\Lambda}_x^{-1}\left(\mathbf{y}_n\right) + \boldsymbol{\Lambda}_x^{-1}\left(\mathbf{y}_m\right)\right)\left(\mathbf{y}_n - \mathbf{y}_m\right) + O\left(\|\mathbf{y}_n - \mathbf{y}_m\|^4\right),$$

where $\boldsymbol{\Lambda}_x\left(\boldsymbol{\psi}_p\right) \triangleq \mathbf{J}_x\mathbf{J}_x^T\left(\boldsymbol{\psi}_p\right) \in \mathbb{R}^{L\times L}$. Since transients are absent for frames $\mathbf{y}_n$ and $\mathbf{y}_m$, the estimated statistics of the observable signal are related to the generating variables of speech, and, by revisiting Appendix III, we have $\mathbf{C}_n^{-1} \approx \boldsymbol{\Lambda}_x^{-1}\left(\mathbf{y}_n\right)$ and $\mathbf{C}_m^{-1} \approx \boldsymbol{\Lambda}_x^{-1}\left(\mathbf{y}_m\right)$. Therefore, (7.6) holds under the hypothesis that only speech is present in both $\mathbf{y}_n$ and $\mathbf{y}_m$. The derivation of (7.6) is analogous in case $\mathbf{y}_n$ and $\mathbf{y}_m$ comprising only transients.

Our derivation of (7.6) concludes by addressing the case where $\mathbf{y}_n$ contains only speech and $\mathbf{y}_m$ contains only transients. In this case, we exploit the introduction of the mid-point, which comprises both speech and transients. As a result, the derivation of (7.13) remains unchanged, and by revisiting Appendix II, we have

$$\|\boldsymbol{\psi}_n - \boldsymbol{\psi}_m\|^2 = \tfrac{1}{2}\left(\mathbf{y}_n - \mathbf{y}_m\right)^T \left(\boldsymbol{\Lambda}_x^{-1}\left(\mathbf{y}_n\right) + \boldsymbol{\Lambda}_t^{-1}\left(\mathbf{y}_m\right)\right)\left(\mathbf{y}_n - \mathbf{y}_m\right) + O\left(\|\mathbf{y}_n - \mathbf{y}_m\|^4\right), \tag{8.5}$$

where $\boldsymbol{\Lambda}_t\left(\boldsymbol{\psi}_m\right) \triangleq \mathbf{J}_t\mathbf{J}_t^T\left(\boldsymbol{\psi}_m\right) \in \mathbb{R}^{L\times L}$. Recall that according to Appendix III, $\mathbf{C}_n^{-1} \approx \boldsymbol{\Lambda}_x^{-1}\left(\mathbf{y}_n\right)$ and $\mathbf{C}_m^{-1} \approx \boldsymbol{\Lambda}_t^{-1}\left(\mathbf{y}_n\right)$. Thus, by substituting them in (8.5), we obtain (7.6).

עיבוד אותות רב-מודאליים על גבי יריעות

דוד דב

עיבוד אותות רב-מודאליים על גבי יריעות

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר

דוקטור לפילוסופיה

דוד דב

הוגש לסנט הטכניון – מכון טכנולוגי לישראל

תמוז תשע"ח חיפה יולי 2018

תודות

המחקר נעשה בהנחיית פרופ' ישראל כהן ופרופ' רונן טלמון מהפקולטה להנדסת חשמל.

אני רוצה להביע את הערכתי למנחים שלי על ההנחיה, ההדרכה והתמיכה במהלך המחקר.

אני רוצה להודות לשיונגקואו מין מאוניברסיטת ג'יאו טונג בשנחאי על כך שסיפק עבור מחקר זה את מערך הנתונים וההקלטות עבור הניסוי של שיערוך מיקום מבט.

אני מודה לטכניון, לקרן ג'ייקובס, לקרן הלאומית למדע (מענקים מס' 576/16 ו-1490/16) ולקרן הלאומית למדע בשיתוף עם הקרן הלאומית למדעי הטבע של סין (מענק מס' 2514/17) על התמיכה הכספית הנדיבה בהשתלמותי.

תקציר

עיבוד אותות רב-מודאליים הינו תחום מחקר המתמקד באנליזה של אותות, אשר נמדדים על ידי חיישנים מרובים מסוג שונה. בשנים האחרונות, תחום זה צובר עניין רב הן בקהילת עיבוד האותות והן בקהילת אנליזת המידע. זאת בשל השימוש הרב בחיישנים שונים במגוון תחומים כגון רפואה, בידור, ביטחון ונהיגה אוטונומית. לדוגמה, מערכים מרובי מיקרופונים ומצלמות וידאו משולבים באופן קבוע במחשבים ניידים וטלפונים חכמים, כמו גם ברכבים אוטונומיים יחד עם חיישנים כגון GPS ומודדי תאוצה.

אותות רב-מודאליים מכילים פעמים רבות מידע עשיר ומגוון: לכל אות בחיישנים השונים, מאפיינים ייחודים כגון ממדים, דינמיקה ותחום ערכים שונה. בפרט, המידע מהחיישנים השונים יכול להיות משותף או ייחודי לכל חיישן. תכונות אלו, יכולות להוות יתרון בשימוש באותות רב-מודאליים ביישומים רבים. אולם, תכונות אלו פותחות צוהר לשאלות מחקריות מהותיות כגון: כיצד למזג אותות רב-מודאליים; כיצד לעבד מידע שזמין בחיישנים השונים רק במקטעי זמן מסוימים; ואיך למדוד באיזה רמה שני אותות שנקלטו מחיישנים מסוג שונה תואמים (קורלטיביים) אחד לשני.

במחקר זה, אנו מטפלים בשאלות פתוחות אלו על ידי פיתוח שיטות גיאומטריות מבוססות גרעין, תוך התבססות על גישת למידה על גבי יריעות (manifold learning). שיטות גרעין קלאסיות בדרך כלל משמשות עבור עיבוד מידע הנקלט על ידי חיישן בודד. הרעיון המרכזי בהן הוא לקבל ייצוגים ממדמד נמוך של האות, מתוך פירוק לווקטורים עצמיים של גרעין דמיון המוגדר על סמך קשרים (אפיניות) בין הדגימות של האות. ייצוגים מסוג זה, בדרך כלל משמרים קשרים מקומיים בין הדגימות, כלומר את הגיאומטריה, והם משמשים בהצלחה במגוון יישומים כגון גילוי אנומליות ומטרות והורדת רעשים בדיבור.

בשנים האחרונות, הוצעו מספר הרחבות לשיטות הגיאומטריות למקרה הרב-מודאלי. הרעיון המרכזי בשיטות אלו, הוא להגדיר גרעין דמיון לאותות בכל חיישן בנפרד, ואז למזג את המידע על ידי שילוב בין הגרעינים. צורת המיזוג שרלוונטית במיוחד למחקר זה, היא מיזוג על ידי מכפלה בין הגרעינים. לשיטה זו, קיימת הפרשנות של דיפוסיה המופעלת על המידע על סמך כל מודאליות בנפרד בצורה מתחלפת. במקרים בהם האות מכיל מספר רב של מקורות, הדיפוסיה על פי מודאליות מסוימת (חיישן מסוים) מפחיתה מקורות המופיעים רק במודאליות האחרת (חיישן אחר). תכונה זו הינה שימושית בתחום עיבוד האותות, מפני שביישומים רבים המידע בו מתעניינים מופיע ביותר מחיישן אחד, בעוד הפרעות הן ייחודיות לסוג החיישן. כך, מיזוג על ידי גרעין המכפלה מקטין את השפעת

ההפרעות. למרות יתרונות אלו, אספקטים חשובים של שיטת מיזוג זו עדיין טרם נחקרו. בפרט, לא ברור כיצד משפיע היחס בין האמפליטודות של המקורות השונים על המיזוג וכיצד על סמך יחס זה, יש להגדיר את גרעיני הדמיון עבור כל חיישן.

תחילתו של מחקר זה, הוא בטיפול בשאלה: כיצד למזג מידע הנקלט מחיישנים רב-מודאליים, כדי לקבל ייצוג של המידע על פי המקורות המשותפים. אנו מציגים ניתוח חדש לשיטת מיזוג המבוססת על מכפלת הגרעינים, תוך התבססות על תורת הגרפים. בפרט, אנו מנתחים את הקשר בין חיבוריות של גרפים המוגדרים על ידי הגרעינים החד והרב-מודאליים. ניתוח זה מאפשר הבנה מעמיקה של השאלה כיצד לבחור את גרעין הדמיון עבור כל חיישן, ובפרט כיצד לבחור את רוחב הגרעין, שהינו פרמטר חשוב השולט בסקלה של הגאומטריה הנלמדת. ככל הידוע לנו מהספרות, הבחירה של פרמטר זה מתבצעת לרוב כמו במקרה של חיישן בודד. אנו מראים שבחירה זו אינה אופטימלית בסביבה של מקורות מפריעים ומציעים אלגוריתם משופר לבחירה נכונה של רוחב הגרעין.

בהמשך, אנו מרחיבים את בעיית מיזוג המידע למקרה של עיבוד מקוון (online), תחת ההנחה הפרקטית שהמידע מהחיישנים השונים זמין רק במקטעי זמן מסוימים. אנו מראים כי ניתן לקבל ייצוג על סמך מקורות משותפים, כלומר להפחית את השפעת ההפרעות, על סמך מקטעי זמן קצרים המכילים מידע מכלל החיישנים. לאחר מכן, אנו מראים כיצד ניתן להשתמש בייצוג זה עבור דגימות חדשות, אפילו אם אלו זמינות רק מחיישן בודד.

אנו מדגימים את השימוש בשיטת המיזוג המשופרת עבור וריאנטים שונים של בעיית גילוי מקורות שמע בהקלטות משולבות אודיו ווידאו. המטרה בבעיה זו, היא לגלות נוכחות מקור הנקלט במשותף במיקרופון ובמצלמת וידאו, תוך התעלמות ממקורות שמע אחרים. לדוגמה, אם מצלמת הווידאו מכוונת אל פניו של דובר מסוים, אז הוא מקור השמע שיש לגלות, תוך התעלמות מדוברים אחרים. אנו מראים כי שיטת המיזוג המוצעת מאפשרת לקבל ייצוג משופר של המקור המשותף, ועל סמך ייצוג זה, להגדיר מדד גילוי, אשר משיג תוצאות משופרות ביחס לאלגוריתמים מתחרים.

בהמשך, אנו מטפלים בשאלה: באיזה מידה אותות מחיישנים מסוג שונה תואמים אחד לשני, במובן שהם חולקים תוכן משותף. אנו מציעים להשתמש בעקבה של מכפלת הגרעינים כמדד התאמה. אנו מראים איך וריאנטים של מדד זה מתקבלים הן מהאנליזה שהצענו המבוססת על תורת הגרפים והן כפרשנות של מכפלת הגרעינים כמשערך של צפיפות של אות בחיישן אחד על סמך חיישן אחר. לבסוף, בהינתן דגימה חדשה, אנו מראים כיצד לעדכן את המדד המוצע בצורה יעילה וללא הצורך לחשבו מחדש. אנחנו מדגימים את השימוש במדד ההתאמה המוצע עובר היישום של זיהוי מקור אודיו בתוך וידאו, וכן עבור שיערוך מיקום מבט של צופה בסרטי וידאו. היישום השני קשור ישירות לראשון, שכן מחקרים בתחום הפסיכולוגיה מראים שאנשים נוטים להסתכל לעבר מקורות אודיו בזמן צפייה בסרטי וידאו. אנו מראים שהמדד המוצע מאפשר לזהות בהצלחה אזורים

בווידאו, להם התאמה גבוהה לאות האודיו, תוך השגת ביצועים עדיפים על פני אלגוריתמים מתחרים.

בחלק האחרון של מחקר זה, אנחנו עוסקים במקרה של אות הנקלט בחיישן בודד בסביבה של הפרעות בלתי רצויות. המטרה שלנו היא לקבל ייצוג של האות, תוך הפרדתו מההפרעות והורדת השפעתן. מנקודת המבט של גישות מבוססות גרעין, המפתח לקבלת ייצוג כזה הוא מציאת מטריקה (וכתוצאה מכך גרעין אפיניות) שמפרידה ביניהם בצורה טובה. בעיה זו מאתגרת במיוחד, לדוגמה, במקרה של אותות דיבור הנקלטים בסביבה של הפרעות טרנזיאנטיות כגון הקשות מקלדת. זאת משום שלעיתים קרובות, אותות הדיבור וההפרעות דומים אחד לשני במובן המטריקה האוקלידית. לכן, ייצוג שנבנה על סמך מטריקה זו אינו מאפשר הפרדה מספקת ביניהם. כדי לטפל בבעיה זו, אנו מציעים מטריקה המבוססת על הסטטיסטיקה של האות בחלונות זמן קצרים ומצדיקים את השימוש בה על סמך הצגה של מודל משתנים חבויים עבור האות וההפרעות. אנו מראים שהמטריקה המוצעת מקרבת את המרחק האוקלידי בין המשתנים החבויים ובפרט מפחיתה את ההשפעה של ההפרעות, תחת ההנחה שקצב השינוי שלהם בזמן מהיר מזה של הדיבור. על ידי שילוב המטריקה המוצעת בשיטה גאומטרית מבוססת גרעין, אנחנו מפתחים מדד לגילוי דיבור אשר משיג ביצועים משופרים ביחס לשיטות מתחרות עבור מגוון רחב של סוגי הפרעות.