Audio-Visual Speech Processing Using Diffusion Maps and The Scattering Transform

David Dov

Audio-Visual Speech Processing Using Diffusion Maps and The Scattering Transform

Research Thesis

As Partial Fulfillment of the Requirements for
the Degree Master of Science in Electrical Engineering

David Dov

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

Cislev 5775 Haifa November 2014

Acknowledgment

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

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

Many thanks to Prof. Ronen Talmon for the collaboration and the great assistance throughout this research.

I would like to thank my family and friends, who supported me throughout this way and in particular volunteered to participate in the recordings of the data set used in this thesis.

Finally, I would like to thank my wife Chen and my son Idan for their endless love and support.

This research was supported by the Israel Science Foundation (grant no.1130/11).

The Generous Financial Help Of The Technion Is Gratefully Acknowledged.

Contents

List of Figures

List of papers

- D. Dov, R. Talmon and I. Cohen, "Audio-Visual Voice Activity Detection Using Diffusion Maps", submitted to IEEE Transactions on Audio, Speech, and Language Processing
- D. Dov and I. Cohen, "Voice Activity Detection In Presence Of Transients Using The Scattering Transform", accepted for publication in Proc. 28th IEEE Convention of the Electrical and Electronics Engineering in Israel, Eilat, Israel

Abstract

Voice activity detection in the presence of highly non-stationary noise and transient interferences is an open problem. State-of-the-art voice activity detectors which are based on statistical models usually assume that the statistics of noise is slowly varying with respect to speech. This assumption does not hold for transient interferences which are short time interruptions, and the performance of these detectors significantly deteriorates. One solution is to incorporate a video signal which is invariant to the acoustic environment. Although several voice activity detectors based on the video signal were recently presented in the literature, merely few detectors which are based on both the audio and the video signals exist in the literature to date. In this thesis, we present two different approaches for voice activity detection. Both approaches incorporate supervised learning procedures and a labeled training data set is considered. In the first approach, we exploit the video signal and present an algorithm for audio-visual voice activity detection. The algorithm comprises a feature extraction procedure, where the features are designed to separate speech from non-speech frames. Diffusion maps is applied separately and similarly to the features of each modality to embed them in a low dimensional representation. Using the new representation, we propose a measure for voice activity which is based on a supervised learning procedure and the variability between adjacent frames in time. The measures of the two modalities are merged to provide voice activity detection based on both the audio and the video signals. Experimental results demonstrate that the incorporation of both audio and video signals is highly beneficial for voice activity detection. In addition, the improved performance of the proposed algorithm compared to state-of-the-art detectors is demonstrated. In the second approach, we focus on the audio signal and propose an algorithm for voice activity detection which is designed to operate in the presence of transients. We propose a continuous measure for voice activity based on the

Support Vector Machines (SVM) classifier. The measure of voice activity is constructed in a features domain, where the features are based on the scattering transform, include noise estimation, and are designed to separate speech and non-speech frames. Experimental results demonstrate that the proposed algorithm outperforms state-of-the-art detectors for different types of background noises, and in particular accurately classifies frames which contain transient interferences.

Notation

$a[n]$	measured audio signal
$a^s[n]$	clean audio signal
$a^d[n]$	background noise
$a^t[n]$	transient
$A(j,k)$	STFT of the measured signal
$A^s(j,k)$	STFT of the clean signal
$A^d(j,k)$	STFT of the background noise
$A^t(j,k)$	STFT of the transient
$A_R(j,k)$	real part of the STFT of the measured signal
$A_I(j,k)$	imaginary part of the STFT of the measured signal
$\mathbf{a}$	audio frame
$\tilde{\mathbf{a}}$	audio feature vector
$\hat{\mathbf{a}}$	diffusion maps of the audio signal
$\mathbf{a}^{MFCC}$	MFCCs
$\mathcal{A}$	audio data set
$\{\mathcal{A}^{tr}, \mathcal{V}^{tr}\}$	training audio-visual data set
$\{\mathcal{A}^{te}, \mathcal{V}^{te}\}$	test audio-visual data set
C	the number of MFCCs
$d(\cdot)$	estimate of the density of the points in the diffusion maps model
$d(\tau)$	background noise (continuous domain)
$D(\cdot,\cdot)$	diffusion distance
$D_{\max}$	maximal diffusion distance between a pair of frames in the training set
$\mathbf{f}$	data point in the diffusion maps model
$\hat{\mathbf{f}}$	diffusion maps

$\mathcal{F}$	set of data points in the diffusion maps model
H	the height of the video frame
$\mathcal{H}_0$	speech absence hypothesis
$\mathcal{H}_1$	speech presence hypothesis
$\mathcal{H}_s$	hypothesis for a stationary signal
$\mathcal{H}_{ns}$	hypothesis for a non-stationary signal
J^A	time window in the construction of the audio features
J^V	time window in the construction of the video features
J	time window in the construction of the voice activity measure
$k(\cdot,\cdot)$	similarity kernel in the diffusion maps model
$k_d(\cdot,\cdot)$	normalized similarity kernel in the diffusion maps model
K	the dimension of diffusion maps
K^f	number of frequency bins
L^S	time window for smoothing the likelihood ratio
L^{US}	time window for the calculation of $P^{US}(\cdot)$
L_t	distance to a hyperplane
$\tilde{L}_t$	normalized distance to a hyperplane
$\tilde{M}$	the dimension of a data point in the diffusion maps model
$\mathbf{M}$	row stochastic Markov matrix in the diffusion maps model
N	number of frames in the audio and the video data sets
N^{te}	number of frames in the test audio-visual data set
N^{tr}	number of frames in the training audio-visual data set
$\tilde{N}$	number of points in the diffusion maps model
N^{lag}	number of lagged frames
$p_r(\cdot;\mathcal{H}_0)$	PDF conditioned on speech absence hypothesis
$p_r(\cdot;\mathcal{H}_1)$	PDF conditioned on speech presence hypothesis
$P(\cdot)$	measure of voice activity
$P^S(\cdot)$	measure of voice activity which is based on GMM
$P^{US}(\cdot)$	measure of voice activity based on the variability between frames
$P^B(\cdot,\cdot)$	bimodal measure of voice activity
P_t	voice activity measure based on SVM classifier

Q	number of wavelets per octave
$S_m(y, \lambda_1, ..., \lambda_m)$	the scattering transform of the noisy signal
T	the length of the window function
$\mathbf{v}$	video frame
$\mathbf{v}^{MV}$	motion vectors
$\tilde{\mathbf{v}}$	video feature vector
$\hat{\mathbf{v}}$	diffusion maps of the video signal
$\mathcal{V}$	video data set
w_Λ	weighting scalar based on noise estimation
w_t	weighting scalar (continuous domain)
W	the width of the video frame
$\mathbf{w}$	weighting vector in the extension of diffusion maps
$y(\tau)$	noisy speech signal (continuous domain)
y_t	time frame of the noisy signal (continuous domain)
$Y(t, \omega)$	STFT of the noisy speech signal (continuous domain)
$\mathbf{y}_t^M(\lambda)$	Mel-scale spectrogram
$\mathbf{y}_t^S$	scattering features
$\mathbf{y}_t$	feature vector (continuous domain)
$x(\tau)$	clean speech signal (continuous domain)
$z(\tau)$	transient interference (continuous domain)
a	parameter that defines a family of diffusion maps
α	parameter that controls the given weight to the two modalities
$\tilde{\gamma}(j, k)$	a posteriori SNR
$\gamma(j, k)$	a posteriori NSSNR
Γ	likelihood ratio based on GMM
$\Gamma_{\max}$	parameter that bounds the likelihood ratio
ϵ	normalization parameter
λ	the central frequency of the band pass filter
$\{\lambda_k\}$	set of eigenvalues of the row stochastic Markov matrix
$\lambda_s(j, k)$	spectral variance of the clean signal
$\lambda_d(j, k)$	spectral variance of noise

$\lambda_t(j,k)$	spectral variance of the transient
$\Lambda_j(k)$	the log of the likelihood ratio
Λ_j	arithmetic mean of the log of the likelihood ratio
$\tilde{\Lambda}_j$	time averaged arithmetic mean of the log of the likelihood ratio
$\tilde{\xi}(j,k)$	a priori SNR
$\xi(j,k)$	a priori NSSNR
ϕ	window function
ϕ'_k	the kth entry of the extended diffusion maps
$\mathbf{\Phi}$	matrix whose columns consist of the eigenvectors and the eigenvalues of $\mathbf{M}$
$\{\boldsymbol{\phi}_k\}$	set of eigenvectors of the row stochastic Markov matrix
$\psi_\lambda(\tau)$	band pass filter
$\Psi_\lambda(\omega)$	Fourier transform of the band pass filter
$\mathbb{1}_{\mathbf{s}}$	speech indicator
$\mathbb{1}_{\mathbf{s}}(t)$	speech indicator (continuous domain)

Abbreviations

AUC	Area Under the Curve
AVVAD	Audio-Visual Voice Activity Detector
DCT	Discrete Cosine Transform
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
IMCRA	Improved Minima Controlled Recursive Averaging
LRT	Likelihood Ratio Test
MCRA	Minima Controlled Recursive Averaging
MFCC	Mel-Frequency Cepstral Coefficients
MFSC	Mel-Frequency Spectral Coefficients
MMSE	Minimum Mean Square Error
NSSNR	Non-Stationary Signal to Noise Ratio
OM-LSA	Optimally Modified Log Spectral Amplitude
PDF	Probability Density Functions
PSD	Power Spectral Density
ROC	Receiver Operating Characteristic
SNR	Signal to Noise Ratio
STFT	Short-Time Fourier Transform
SVM	Support Vector Machines
VAD	Voice Activity Detector

Chapter 1

Introduction

1.1 Motivation and Overview

Voice activity detection is an essential component in many applications such as speech and speaker recognition [2], speech coding, speech enhancement and dominant speaker identification [3]. Early methods for voice activity detection are based on straightforward features such as the energy of the signal and zero-crossing rate [4]. Although these methods perform well for clean signals, their performances deteriorate in the presence of background noise since for example the zero-crossing rate may be increased due to the background noise, wrongly indicating voice activity. To overcome this problem, several Voice Activity Detectors (VAD) which assume statistical models for the input signal and the noise, and are based on Likelyhood Ratio Test (LRT) were presented in recent years. Among the different statistical models are the Gaussian model [5, 6], a Laplacian model [7–9] and a generalized Gamma model [10]. The LRT, in these methods, usually incorporates an estimation of the spectral variance of the noise, and common methods for its estimation, such as the Minima Controlled Recursive Averaging (MCRA) method [11–13], typically assume that the statistics of noise is slowly varying with respect to speech. This assumption does not hold for highly non-stationary noise and transient interferences, which are short time interruptions such as keyboard typing and door knocking [14, 15], and the performance of these methods significantly deteriorate.

Recently, A supervised learning algorithm for voice activity detection in the presence of highly non-stationary noise was presented in [16]. The algorithm comprises features

selection procedure, where the feature are based on the Mel-Frequency Cepstral Coefficients (MFCC) and noise estimation. Then, a spectral clustering method is applied in the features domain for the classification process. Although the algorithm was shown to outperform several state-of-the-art methods, its performance is still limited in the presence of transients since the transients and speech are only partially separated in the features domain. Therefore, voice activity detection in presence of highly non-stationary noise and transients remains an open problem.

One solution could be the incorporation of visual information which is invariant to the acoustic environment and is available in a variety of speech related products, e.g., teleconference rooms, free internet services of video telephony, and smart phone video callings. Thus, exploiting visual data of speech can be beneficial and practical for voice activity detection in a challenging acoustic environment, and indeed, several algorithms for voice activity detection which utilize the video signal were presented in the literature in recent years. While VADs which are based on the video signal have an advantage over VADs which are based on the audio signal in noisy environments, they fail to compete with state-of-the-art audio VADs in quiet environments, since even straightforward features, such as the energy of the audio signal, are near perfect indicators for speech presence in a quiet environment. Therefore, the incorporation of the audio and the video signals may be advantageous for voice activity detection. Yet, only few algorithms that combine the audio and the video signal for voice activity detection have been presented in the literature to date.

In this thesis we present two supervised learning algorithms for voice activity detection. Both algorithms consider training data sets which contain speech utterances contaminated with background noise and transients, and are designed to perform in a highly non-stationary environment. In the first algorithm, we assume that the speech signal is captured by two different sensors, a single microphone and a video camera, where the video signal comprises a bounding box of the region of the mouth. The algorithm is based on an assumption that speech has an underlying geometric structure which stems from the physics of its production and does not depend on the sensors. We apply diffusion maps, which is a data driven geometric method, similarly and separately to each modality, to reveal the underlying structure. The diffusion maps is applied in a domain

of features which are designed to separate speech and non-speech frames, and provides a low dimensional representation of the signals. The low dimensional representation is used to define a measure for voice activity which incorporates a supervised learning procedure based on Gaussian Mixture Model (GMM) and an unsupervised learning procedure that assumes different variability rates of the signal during speech and non-speech intervals. The measures of voice activity based on the audio and the video signals are finally merged to provide a bimodal measure of voice activity which in turn used for voice activity detection. The performance of the proposed algorithm are evaluated for different types of noises and transients, and compared to state-of-the-art VADs. Experimental results demonstrate improved performance of the proposed algorithm compared to state-of-the-art VADs, and in particular, the advantage of the incorporation of both the audio and the video signals for voice activity detection is demonstrated.

The second algorithm exploits only the audio signal and is based on the scattering transform. The scattering transform is constructed through a set of convolutions of the signal with wavelet filter banks and modulus operators, and provides a representation of the signal in the form of a deep network. High orders of the scattering transform successfully characterize short time scale cues and specifically transients. The proposed algorithm exploits features which are based on the scattering transform, and allows for a good separation between speech and transients, and in addition incorporates noise estimation procedure that allows to separate speech from background noise. The features are used to define a continuous measure of voice activity based on the SVM classifier. Experimental results demonstrate that the proposed algorithm outperforms state-of-the-art detectors and in particular provides accurate detections in frames which contain transients.

1.2 Thesis Structure

The rest of the thesis is organized as follows: In Chapter 2 we survey recent algorithms for voice activity detection based on audio and video signals. In addition, we provide a short theoretical background on Diffusion Maps, which is a manifold learning method and the Scattering transform. In Chapter 3 we present an algorithm for audio-visual voice activity detection using diffusion maps and evaluate its performance. In Chapter 4 we present an

algorithm for voice activity detection which exploits only the audio signal and is based on the scattering transform. In Chapter 5 we conclude the findings of this research and discuss further research directions.

Chapter 2

Related Work and Theoretical Background

2.1 Introduction

Voice activity detection in a noisy environment is an open problem which was extensively studied in the last two decade. In Section 2.2 we review milestone and recent methods for voice activity detection based on the audio signal and show their dependency on accurate estimation of noise. In Section 2.3 we review a well known method for noise estimation and explain its limitations in the presence of transients, which are short term interruptions, e.g., keyboard typing, door knocking and office noise [14, 15]. In this thesis we propose to incorporate the video signal which is invariant to the acoustic environment. In Section 2.4 we review methods for voice activity detection which are based on the video signal and remark that only few studies deal with the incorporation of the audio and the video signals. In this thesis the representation of the audio and the video signals is based on diffusion maps, which is a geometric method for dimensionality reduction, and is reviewed in Section 2.5. In addition, we exploit the scattering transform for the representation of audio signals, and we provide a brief theoretical background for the scattering transform in Section 2.6.

2.2 Audio Voice Activity Detection

Early methods for voice activity detection which are based on the audio signal, such as the G.729 standard [4], are based on straightforward features such as the energy of the signal and zero-crossing rate. Although they perform well for clean signals, the performances of these methods deteriorate in the presence of background noise since for example the zero-crossing rate may be increased due to background noise, wrongly indicating on voice activity.

To overcome this problem, several VADs which assume statistical models to the input signal and the noise, and are based on LRT were presented in the last two decades [5–10]. Next, we briefly review the general structure of these methods and the related statistical models. Let $a[n]$ be the measured speech signal contaminated with additive background noise, given by:

$$a[n] = a^s[n] + a^d[n] \tag{2.1}$$

where $a^s[n]$, and $a^d[n]$ are speech and the background noise, respectively. We remark that in this section, we assume a traditional model of the noisy signal which includes the speech signal and the background noise but not transients, and the effect of transients is discussed in Section 2.3. The signal is processed using consecutive frames in the Short-Time Fourier Transform (STFT) domain, such that equation (2.1) is given in the STFT domain by:

$$A(j,k) = A^s(j,k) + A^d(j,k) \tag{2.2}$$

where $A(j,k)$, $A^s(j,k)$ and $A^d(j,k)$ are the STFTs of $a[n]$, $a^s[n]$ and $a^d[n]$, respectively, $k = 1, 2, ...K^f$ is the frequency bin index, and $j = 1, 2, ...$ is the time frame index. Consider two hypotheses, $\mathcal{H}_0$ and $\mathcal{H}_1$, denoting speech absence and speech presence, respectively, and let $p_r(\cdot;\mathcal{H}_0)$ and $p_r(\cdot;\mathcal{H}_1)$ be Probability Density Functions (PDF) of the input signal in the STFT domain conditioned on the hypotheses. Accordingly, a log of the likelihood ratio is defined:

$$\Lambda_j(k) = \log\left(\frac{p_r\left(A(j,k);\mathcal{H}_1\right)}{p_r\left(A(j,k);\mathcal{H}_0\right)}\right). \tag{2.3}$$

In the method presented in [5], the signal and the noise are assumed to be complex Gaussian uncorrelated processes in the STFT domain, i.e., the conditional PDFs are given

by:

$$p_r\left(A(j,k);\mathcal{H}_0\right) = \frac{1}{\pi\lambda_d(j,k)}\exp\left\{-\frac{|A(j,k)|^2}{\lambda_d(j,k)}\right\}, \tag{2.4}$$

$$p_r\left(A(j,k);\mathcal{H}_1\right) = \frac{1}{\pi\left[\lambda_s(j,k)+\lambda_d(j,k)\right]}\exp\left\{-\frac{|A(j,k)|^2}{\lambda_s(j,k)+\lambda_d(j,k)}\right\}, \tag{2.5}$$

where $\lambda_s(j,k)$ and $\lambda_d(j,k)$ are the spectral variance of speech and noise, respectively. For the Gaussian model, it can be shown that $\Lambda_j(k)$ is given by [5, 17]:

$$\Lambda_j(k) = \left(\frac{\tilde{\gamma}(j,k)\cdot\tilde{\xi}(j,k)}{1+\tilde{\xi}(j,k)} - \log\left(1+\tilde{\xi}(j,k)\right)\right), \tag{2.6}$$

where $\tilde{\xi}(j,k) \triangleq \frac{\lambda_s(j,k)}{\lambda_d(j,k)}$ and $\tilde{\gamma}(j,k) = \frac{|A(j,k)|^2}{\lambda_d(j,k)}$ are called a priori and a posteriori SNR, respectively. We remark that both the a priori SNR, $\tilde{\xi}(j,k)$, and the spectral variance of the noise, $\lambda_d(j,k)$, are also exploited in the following reviewed methods. The estimation of the spectral variance of the noise from the noisy signal is reviewed in Section 2.3, and in the following we review the estimation of the a priori SNR using the Decision Directed Method [17]. Let $\hat{\lambda}_d(j,k)$ be the estimated spectral variance, and let $\hat{A}_s(j,k)$ be the Minimum Mean Square Error (MMSE) estimator of the speech signal, the estimated a priori SNR is given by:

$$\hat{\xi}(j,k) = \alpha\frac{|\hat{A}_s(j-1,k)|^2}{\hat{\lambda}_d(j-1,k)} + (1-\alpha)\tilde{P}[\hat{\gamma}(j,k)-1],$$

where $\alpha \in (0,1)$ is a smoothing parameter, $\tilde{\gamma}(j,k)-1$ is the maximum likelihood estimator of $\tilde{\xi}(j,k)$, and $\tilde{P}(x) = \left\{\begin{array}{ccc} x & ; & x \geq 0 \\ 0 & ; & \text{otherwise} \end{array}\right\}$ and is used for keeping non-negative values of the estimator [17].

The estimated log of the likelihood ratio in (2.6) is averaged over the frequency bins, and the arithmetic mean of the log of the likelihood ratio is given by:

$$\Lambda_j = \frac{1}{K^f}\sum_{k=1}^{K^f}\Lambda_j(k). \tag{2.7}$$

Λ_j is a scalar feature that indicates on voice activity in frame index j, such that the higher Λ_j the higher is the probability that speech is present in the frame. Voice activity

is detected in [5], by comparing Λ_j to a threshold such that speech is present in the frame if Λ_j is greater than a certain threshold. Finally the decision of the VAD is smoothed using a Hang-Over scheme which assumes a strong correlation in the occurrences of consecutive speech frames [5].

The authors in [6] extended the method presented in [5] by assuming that long-term speech information can be beneficial for voice activity detection due to high correlation between consecutive speech frames. They proposed a VAD which is based on an LRT that involves multiple observations, such that the log of the likelihood ratio based on several observations, denoted by $\tilde{\Lambda}_j(k)$, is given by:

$$\tilde{\Lambda}_j(k) = \log\left(\frac{p_r\left([A(j-J,k), A(j-J+1,k), ..., A(j+J,k)]\,;\mathcal{H}_1\right)}{p_r\left([A(j-J,k), A(j-J+1,k), ..., A(j+J,k)]\,;\mathcal{H}_o\right)}\right),$$

where $[A(j-J,k), A(j-J+1,k), ..., A(j+J,k)]$ is a row vector of $2J+1$ observations. Assuming that the observations are independent, it is shown in [6] that $\tilde{\Lambda}_j(k)$ is given by:

$$\tilde{\Lambda}_j(k) = \frac{1}{2J+1}\sum_{i=j-J}^{j+J}\Lambda_i, \tag{2.8}$$

where Λ_i is the arithmetic mean of the log of the likelihood ratio based a single observation, defined in (2.7). The authors reported on superior performance using the multiple observations approach compared to the approach based on a single observation.

In the studies presented in [7, 8] and [9] a Laplacian distribution is assumed for the STFTs of the signals. According to this assumption, the conditional PDFs of the input signal are given by:

$$p_r(A(j,k);\mathcal{H}_0) = \frac{1}{\lambda_s(j,k)+\lambda_d(j,k)}\exp\left\{-\frac{2\left(|A_R(j,k)|+|A_I(j,k)|\right)}{\sqrt{\lambda_s(j,k)}}\right\},$$

$$p_r(A(j,k);\mathcal{H}_1) = \frac{1}{\lambda_s(j,k)+\lambda_d(j,k)}\exp\left\{-\frac{2\left(|A_R(j,k)|+|A_I(j,k)|\right)}{\sqrt{\lambda_s(j,k)+\lambda_d(j,k)}}\right\},$$

where $A_R(j,k)$ and $A_I(j,k)$ are the real and the imaginary parts of the input signal, respectively, and are assumed to be independent. Accordingly, the log of the likelihood

ratio based on the Laplacian model is given by:

$$\Lambda_j(k) = 2\left(|A_R(j,k)| + |A_I(j,k)|\right)\left(\frac{|A(j,k)| - \sqrt{\lambda_d(j,k)}}{\sqrt{|A(j,k)|\lambda_d(j,k)}}\right) - \log\left(1 + \tilde{\xi}(j,k)\right). \quad (2.9)$$

It is empirically shown by the authors in [8] that the Laplacian distribution better fits the spectrum of the noisy signal compared to the Gaussian distribution for different types of noises, including vehicular noise and babble noise. In addition, it was reported that the LRT based on the Laplacian model provides better classification results with respect to the Gaussian model. In the method presented in [9], the LRT is further extended to incorporate the correlation between consecutive frames. The likelihood ratio is based on PDFs of the signal which are conditioned not only on the hypotheses of speech presence and absence in the current frame, but also on the hypotheses in the previous one.

In the study presented in [18], the authors tested the fitness of Gamma distribution in addition to Gaussian and Laplacian distributions to the spectrum of the noisy signal. The PDFs of the Gamma distribution conditioned on the two hypotheses are given by:

$$p_r(A(j,k);\mathcal{H}_0) = \frac{\sqrt{6}}{8\pi\sqrt{\lambda_d(j,k)|A_R(j,k)A_I(j,k)|}}\exp\left\{-\frac{\sqrt{3}\left(|A_R(j,k)| + |A_I(j,k)|\right)}{\sqrt{2}\sqrt{\lambda_d(j,k)}}\right\},$$

and:

$$p_r(A(j,k);\mathcal{H}_1) = \frac{\sqrt{6}}{8\pi\sqrt{\left(\lambda_s(j,k) + \lambda_d(j,k)\right)|A_R(j,k)A_I(j,k)|}}\exp\left\{-\frac{\sqrt{3}\left(|A_R(j,k)| + |A_I(j,k)|\right)}{\sqrt{2}\sqrt{\lambda_s(j,k) + \lambda_d(j,k)}}\right\}.$$

It is found that the fitness of a particular model to the spectrum of the noisy signal depends on the type of the noise and the SNR, and a method for voice activity detection is presented, where the statistical model is adaptively chosen in an online manner according to the most fitting model.

Recently, a novel VAD which is designed to perform in a highly non-stationary environments was presented in [16]. The method is based on a supervised learning procedure and exploits features which are based on MFCCs. The MFCCs provide a time frequency representation of the signal which is widely used for speech recognition and its construction is described in Section 2.6.1. Unlike previous methods which are entirely based on

the likelihood ratio, in this method the arithmetic mean of the log of the likelihood ratio, Λ_j in (2.7), is used in the features level to weight the MFCCs, such that low weights are assigned to frames which contain only background noise. The classification to speech and non-speech frames is given by a spectral clustering method and improved performance of voice activity detection in the presence of highly non stationary noise is reported.

2.3 Noise Spectrum Estimation and The Effect of Transients

Common to the methods described throughout the previous section is that they all rely on the estimation of the spectral variance of the noise, $\lambda_d(j,k)$, and their performance highly depends on accurate estimation of noise statistics. One of the well known and widely used methods for the estimation of the spectral variance of the noise is the MCRA method [11–13], where the spectral variance is recursively smoothed in speech absence intervals. Although this method have some extensions such as the method in [19], for simplicity, we review the early method. In addition, despite the success of this method in the estimation of the spectral variance of various quasi-stationary types of noises, we explain in this section why it is not adequate for the estimation of the spectral variance of transients.

The estimate of the spectral variance of the noise is given by smoothing the energy of the noisy signal over time in speech absent intervals and keeping previous values of the estimate in speech intervals. More specifically:

$$\hat{\lambda}_d(j+1,k) = \left\{ \begin{array}{ccc} \alpha_d \hat{\lambda}_d(j,k) + (1-\alpha_d)|A(j,k)|^2 & ; & \mathcal{H}_0' \\ \hat{\lambda}_d(j,k) & ; & \mathcal{H}_1' \end{array} \right\}, \tag{2.10}$$

where $\alpha_d \in (0,1)$ is a smoothing parameter, and $\mathcal{H}_0'$ and $\mathcal{H}_1'$ are speech absence and speech presence hypotheses for noise spectrum estimation, respectively. These hypotheses are distinct from $\mathcal{H}_0$ and $\mathcal{H}_1$ such that a decision that speech is absent is taken with a lower confidence, i.e., $p_r(\mathcal{H}_0'|A(j,k)) \leq p_r(\mathcal{H}_0|A(j,k))$. The hypotheses $\mathcal{H}_0'$ and $\mathcal{H}_1'$ are used in order to avoid the smoothing of $\hat{\lambda}_d$ over speech components due to wrong speech detections. Since $\mathcal{H}_0'$ and $\mathcal{H}_1'$ are unknown and can be estimated up to some degree of

certainty, (2.10) implies:

$$\hat{\lambda}_d(j+1,k) = \hat{\lambda}_d(j,k)p'(j,k) + \left[\alpha_d\hat{\lambda}_d(j,k) + (1-\alpha_d)|A(j,k)|^2\right](1-p'(j,k)),$$

where $p'(j,k)$ is the probability of speech presence conditioned on the noisy signal, i.e., $p'(j,k) = p_r(\mathcal{H}_1'|A(j,k)$.

$p'(j,k)$ is estimated by tracking the energy of the noisy signal which is smoothed in time and frequency. Let $S_f(j,k)$ be a local energy of the signal smoothed in the frequency domain using a window of length $2\tilde{w}+1$:

$$S_f(j,k) = \sum_{i=-\tilde{w}}^{\tilde{w}} b(i)|A(j,k-i)|^2,$$

where $b(i)$ is a window function. The energy of the noisy signal, which is smoothed in time and frequency, is denoted by $S(j,k)$, and is given by smoothing $S_f(j,k)$ over time:

$$S(j,k) = \alpha_s S(j-1,k) + (1-\alpha_S)S_f(j,k),$$

where $\alpha_s \in (0,1)$ is a smoothing parameter. Speech is assumed to be present, i.e., the hypothesis $\mathcal{H}_1'$ is assumed to hold, when the smoothed energy of the signal is greater than the energy of the noise. Let $I(j,k)$ be an indicator of speech, which is given by:

$$I(j,k) = \left\{ \begin{array}{lcc} 1 & ; & S(j,k) > \delta S_{\min}(j,k) \\ 0 & ; & \text{otherwise} \end{array} \right\}, \tag{2.11}$$

where δ is a constant parameter, and $S_{\min}(j,k)$ is the minimum of the smoothed energy of the noisy signal in a causal temporal neighborhood and is associated with the energy of the noise. The estimate of the conditioned speech presence probability, $p'(j,k)$, is given by taking into account the strong correlation between consecutive speech frames, and smoothing the indicator $I(j,k)$ in (2.11) over time to reduce frequent fluctuations between speech presence and speech absence decisions. In particular, $p'(j,k)$ is given by:

$$p'(j,k) = \alpha_p p'(j-1,k) + (1-\alpha_p)I(j,k),$$

where $\alpha_p \in (0, 1)$.

We remark that the main assumption in the estimation of the spectral variance of the noise lies in equation (2.11), where speech is associated with abrupt changes of the smoothed energy of the signal. Namely, the noise is assumed to slowly vary with respect to speech. This assumption does not hold for transient interferences which are characterized by abrupt bursts of sound and vary faster than speech [14, 15, 20]. Therefore, transients are wrongly estimated as speech by the indicator of speech presence, $I(j, k)$, leading to false low levels of the spectral variance of the noise and to high levels of the log of the likelihood ratio. As a result, the performance of VADs which are based on the LRT significantly deteriorates in the presence of transients. In the method presented in [16], the arithmetic mean of the log likelihood ratio, Λ_j, is used to weight the features. In the presence of transients, high weights are assigned both in presence of speech and transient which are the fast varying components of the measured signal. Therefore, the features only partially separate speech from the transients, and the performance of the VAD is still limited. Voice activity detection in presence of transients remains an open problem.

2.4 Visual Voice Activity Detection

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

Existing voice activity detection methods, which are based on visual data, focus on the analysis of the region of the mouth, and in particular the lips. However, their main drawback is their dependency on the detection of the lips, which often rely on artificial markers and whose accuracy may be degraded due to skin color or illumination conditions. For example, in the studies presented in [21] and [22], the detection is based on features, which are constructed from contours of the lips. However, the extraction of the contours requires that the lips would be marked using a blue makeup. In [23], the presented VAD exploits the shape and color of the lips, which are obtained assuming that the lips are

marked by key points. Another approach for extracting the lips, which assumes that the color of lips is significantly different from the color of skin, was proposed in [24] and [25].

Another approach to visual voice activity detection relies on the dynamics of the region of the mouth. In [23], a second algorithm based on the movements of lips was presented as well. This algorithm focuses on the analysis of the region of the mouth, which is enhanced using a retinal filter. Although exhibiting good performance, the detection was found to be sensitive to lips movements in speech absent intervals. In [26], a similar approach based on motion estimation in the region of mouth was presented. Motion fields are used as features to characterize the change of the position of the mouth over time and a Hidden Markov Model (HMM) is used for the classification. Such a motion estimation approach was also utilized in [27]. There, the energy in the mouth region is defined using optical flow and serves as a feature for a classifier based on an HMM. A VAD based on intensity values in the region of the mouth was presented in [28], where the detection is based on the number of low intensity pixels, which are modeled using a Gaussian model for speech and non-speech hypotheses. Although high detection rates were reported, the algorithm may be limited in real time applications because the entire speech sequence is required in advance to estimate the noise statistics.

Although VADs based on video signals have an advantage over VADs based on audio signals in noisy conditions, they usually fail to compete with audio based detectors, since a trivial classifier based on the energy of the audio signal can obtain near perfect performance in quiet environments. Therefore, a bimodal VAD may combine the advantages of both audio and video signals. Yet, only few detectors based on both audio and video signals exist in the literature to date. Such an Audio-Visual Voice Activity Detector (AVVAD) was presented in [29], where the video signal is analyzed using a Bayesian approach to detect the lips, followed by an HMM to model the lips movements. The audio signal, which is assumed to be acquired in a microphone array, is used to compute a spatio-temporal coherence of the source. Then, another HMM is used for speech presence estimation. Finally, the two modalities are combined at the classification stage using a tree based classifier.

2.5 Diffusion Maps

In recent years, great efforts were made to develop tools for the analysis of high dimensional data, based on its geometric structure. The main assumption of these methods, which are also referred to as manifold learning, is that the observable high dimensional data lies on a low dimensional manifold. Namely, the data has an intrinsic compact geometric structure, and as a result, can be represented in a compact way. The goal of these methods is to reveal the underlying structure of the data, and provide a parametrization of the data on the manifold, i.e., embed (map) the data to a low dimensional Euclidean domain. In the methods presented in [30–34] a Markov chain is constructed on a graph of the data, and the main idea is that eigenvectors of Markov matrices may be seen as the underlying coordinates of the data. Diffusion maps, is a general framework for manifold learning, where the aforementioned methods may be seen as a special case. Next we briefly describe the construction of diffusion maps.

Let $\mathcal{F} = \{\mathbf{f}_i\}_{i=1}^{\tilde{N}}$ be a set of $\tilde{N}$ data points in an $\tilde{M}$ dimensional Euclidean domain. A pairwise similarity kernel function $k(\mathbf{f}_i, \mathbf{f}_j)$ between the ith and the jth data points is defined as:

$$k(\mathbf{f}_i, \mathbf{f}_j) = e^{-\frac{||\mathbf{f}_i - \mathbf{f}_j||^2}{\varepsilon}}, \tag{2.12}$$

where $||\cdot||$ is the L_2 norm and ε is the kernel bandwidth chosen according to [35]. Since the data points are not uniformly distributed on the manifold, the kernel is normalized by an estimate of the density of the data points on the manifold [36]:

$$k_d(\mathbf{f}_i, \mathbf{f}_j) = \frac{k(\mathbf{f}_i, \mathbf{f}_j)}{(d(\mathbf{f}_i))^a (d(\mathbf{f}_j))^a}, \tag{2.13}$$

where $d(\mathbf{f}_i)$ is the estimate of the density given by:

$$d(\mathbf{f}_i) = \sum_{\mathbf{f}_j \in \mathcal{F}} k(\mathbf{f}_i, \mathbf{f}_j). \tag{2.14}$$

The parameter $a \in [0, 1]$ defines a family of diffusion maps, and controls the degree of the influence of the density of the points on the embedding of the data. For the special case where $a = 0$, the diffusion maps are similar to the classical manifold learning

methods [32, 34] where the influence of the density is maximal. For the other extreme case, where $a = 1$, diffusion maps provides density invariant embedding and recover the the Riemannian geometry of the manifold, even if the data points are not distributed uniformly on the manifold [36]. This case is in particular adequate for merging data captured by different types of sensors (in this study a microphone and a video camera), with different densities [35, 37]. Based on the normalized kernel, a weighted symmetric graph is constructed, where each data point, $\mathbf{f}_i$, is viewed as a node and the weight of the edge connecting nodes i and j is given by $k_d(\mathbf{f}_i, \mathbf{f}_j)$. A Markov chain on the graph is defined by normalizing the kernel once again. Let $\mathbf{M} \in \mathbb{R}^{\tilde{N}\times\tilde{N}}$ be a row stochastic Markov matrix, given by:

$$\mathbf{M}_{i,j} = \frac{k_d(\mathbf{f}_i, \mathbf{f}_j)}{s(\mathbf{f}_j)}, \tag{2.15}$$

where

$$s(\mathbf{f}_j) = \sum_{\mathbf{f}_j \in \mathcal{F}} k_d(\mathbf{f}_i, \mathbf{f}_j). \tag{2.16}$$

As a result, the nodes of the graph, $\mathbf{f}_j \in \mathcal{F}$, may be seen as the states of a Markov chain with the transition probability matrix $\mathbf{M}$. Finally, an eigenvalue decomposition is applied to $\mathbf{M}$, yielding eigenvalues $\{\lambda_k\}$, which are sorted in descending order, and corresponding eigenvectors $\{\boldsymbol{\phi}_k\}$. The eigenvalues of $\mathbf{M}$ are in the range of $0 \div 1$ due to the row normalization [36]. Moreover, $\lambda_0 = 1$ and its associated eigenvector $\boldsymbol{\phi}_0$ is an all-ones vector. This constant eigenvector is ignored since it does not contain any information [37].

The first K largest eigenvalues (excluding the trivial) and the corresponding K eigenvectors of $\mathbf{M}$ are used for the parametrization of the data points on the manifold. A matrix $\boldsymbol{\Phi} \in \mathbb{R}^{\tilde{N}\times K}$, whose columns consist of the eigenvectors and the eigenvalues of the transition probability matrix, is formed:

$$\boldsymbol{\Phi} \equiv (\lambda_1\boldsymbol{\phi}_1, \lambda_2\boldsymbol{\phi}_2, ..., \lambda_K\boldsymbol{\phi}_K). \tag{2.17}$$

From (2.17), the diffusion mapping of the feature vector $\mathbf{f}_i$ is given by the ith row of the matrix $\boldsymbol{\Phi}$:

$$\hat{\mathbf{f}}_i = (\boldsymbol{\Phi}_{i,1}, \boldsymbol{\Phi}_{i,2}, ..., \boldsymbol{\Phi}_{i,K}). \tag{2.18}$$

Thus, an embedding $\hat{\mathbf{f}}_i$ of each data point i into a K dimensional Euclidean space is obtained. Due to the assumption, that there exists a low intrinsic structure of the data, the spectrum of the transition probability matrix (the eigenvalues) decays fast. Therefore, entries in (2.17), that correspond to small eigenvalues, are negligible and K may be set to a small value, thereby providing significant dimensionality reduction.

Diffusion maps were previously exploited for audio-visual speech recognition in [35]. The authors proposed to exploit density invariant embedding, i.e., setting $a = 1$ in (2.13), and apply diffusion maps in a similar way to each modality. The data is merged by a concatenation of the diffusion maps of each modality into a super-vector [35, 37], and the superiority of bimodal processing with respect to single modal processing was reported. In addition, improved recognition rates were achieved for density invariant embedding ($a = 1$) with respect to the case where the embedding is influenced by the density of the data on the manifold.

2.6 The scattering transform and comparison to MFCC

In this section we briefly review the MFCC representation of speech signals and the scattering transform and discuss the connection between them. We remark that following [1], we present these two representation in the continuous domain.

2.6.1 MFCC

Let $y(\tau)$ be a (noisy) speech signal and let $Y(t, \omega)$ be the representation of $y(\tau)$ in the STFT domain, given by:

$$Y(t, \omega) = \int y(u)\phi(u - t)e^{-j\omega u}du \tag{2.19}$$

where ϕ is a window function centered at time t. A Mel-frequency spectrogram is defined using Mel-scale filters $\{\psi_\lambda(\tau)\}_\lambda$, where λ is the center frequency, and the Fourier transform of filter $\psi_\lambda(\tau)$ is given by $\Psi_\lambda(\omega)$. The Mel-frequency spectrogram is given by averaging

the energy of the STFT of the signal with Mel-scale filters:

$$\mathbf{y}_t^M(\lambda) = \frac{1}{2\pi} \int |Y(t,\omega)|^2 |\Psi_\lambda(\omega)|^2 d\omega. \tag{2.20}$$

According to the Mel-frequency scale, $\{\Psi_\lambda(\omega)\}_\lambda$ is a filter bank of band pass filters which covers the whole frequency axis. Filters that are centered at high frequencies, have a constant-Q bandwidth of λ/Q, and those that are centered at the low frequencies, have a bandwidth of $2\pi/T$, where T is the length of the time window. The constant-Q filters (for the high frequencies) are given by a complex analytical wavelet $\psi(\tau)$, whose Fourier transform $\Psi(\omega)$ approximately equals zero for $\omega < 0$ and $\Psi(0) = 0$. In practice, all wavelets used in this thesis are Morlet wavelets as in [1]. The wavelet filter bank, $\{\psi_\lambda(\tau)\}_\lambda$, is given by dilating the wavelet, such that $\psi_\lambda(t) = \lambda\psi(\lambda t)$. For the high frequencies, i.e., where $\lambda \geq 2\pi Q/T$, λ is given by $\lambda = 2^{j/Q}, j \in \mathbb{Z}$, where Q is the number of wavelets per octave. For $\lambda < 2\pi Q/T$, $Q-1$ equally spaced filters are used with constant bandwidth $2\pi/T$ to cover the low frequency interval. For simplicity, the lower frequency filters are still termed wavelets. The MFCCs are given by applying a cosine transform on the coefficients of the log of the spectrogram, which decorrelates the coefficients.

2.6.2 The Scattering Transform

The scattering transform may be seen as an extension of the Mel-frequency spectrogram. For large values of λ, $\phi(t)$ is approximately constant on the support of $\psi_\lambda(\tau)$ and it is shown in [1] that the Mel-frequency spectrogram in (2.20) can be approximated in this case by:

$$\mathbf{y}_t^M(\lambda) \approx |y \star \psi_\lambda|^2 \star |\phi|^2(t). \tag{2.21}$$

Thus, the spectrogram is approximately equals to time averaging of $|y \star \psi_\lambda|^2$, where ϕ may be seen as a low pass filter. The MFCCs are efficient descriptors for the representation of speech signals in the time-frequency domain and are widely used in speech recognition [38]. However, averaging frequency bands over the Mel-scale causes a loss of information [1]. For example, the representation of transients, which are short term cues, may be similar

to the representation of speech due to the time averaging.

The scattering transform extends the MFCCs by providing a representation which is more suitable for the representation of short term cues as the lost information is recovered with multiple layers of wavelet coefficients. The scattering transform is defined by cascading wavelets decompositions and modulus operators providing a deep scattering network. Let $S_m(y, \lambda_1, ..., \lambda_m)$ be the scattering transform of order m. The zero order scattering transform is given by $S_0(t) = y \star \phi(t)$, and it approximately equals zero as speech signals contain low energy in low frequencies. Thus, the zero order transform is not of a great importance and it is given here for the sake of completeness. The first order scattering transform approximates the Mel-frequency spectrogram, and similarly to (2.21), is given by:

$$S_1(t, \lambda_1) = |y \star \psi_{\lambda_1}| \star \phi(t) \tag{2.22}$$

where the filter bank, $\{\psi_{\lambda_1}\}_{\lambda_1}$ is similar to the one used in the construction of the MFCCs and ϕ is a low pass filter. Note that the square in (2.21) is removed to reduce the dynamical range. Similarly, the mth order scattering transform is given by:

$$S_m(t, \lambda_1, ..., \lambda_m) = |||y \star \psi_{\lambda_1}| \star ...| \star \psi_{\lambda_m}| \star \phi(t) \tag{2.23}$$

Note that large values of m impose high computational load. In this thesis we use only the first and the seconds order scattering transforms as higher order transforms have not found to provide significant advantage in previous studies [1] and [39]. The scattering network which is defined by (2.23) is illustrated in Fig. 2.1.

2.7 Discussion

In this chapter we reviewed state-of-the-art methods for voice activity detection. Methods which are based on the audio signal, typically rely on the estimation of the spectral variance of the noise. The latter is estimated based on the basic assumption that speech is abruptly varying, and therefore, the noise is characterized as the slowly varying part of the noisy signal. However, transients are also abruptly varying over time and are not estimated as the non-speech part of the signal. As a result, transients are wrongly

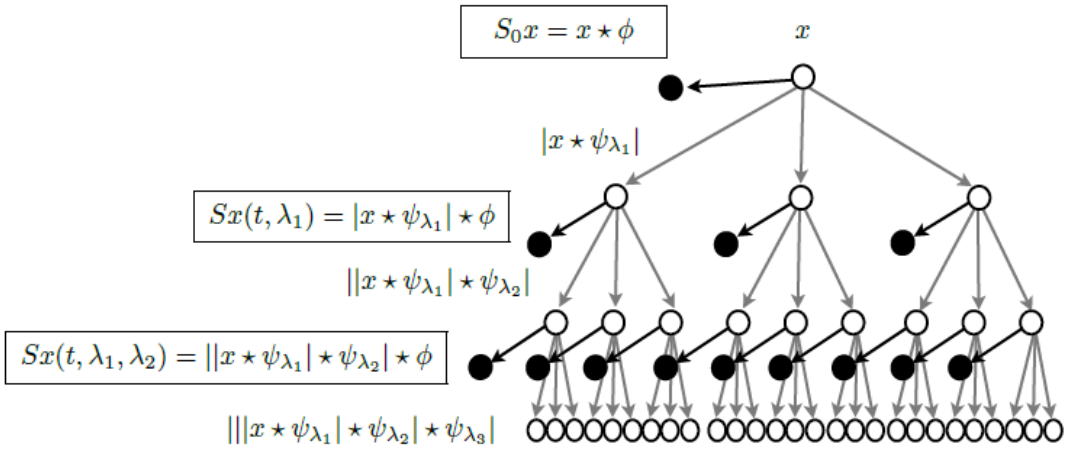


Figure 2.1: The scattering network which is given by applying a cascade of wavelets and modulus operators (Image courtesy [1]).

detected as speech and the performance of state-of-the-art VADs significantly deteriorated. In this study we propose to incorporate the video signal which is invariant to the acoustic environment and can be beneficial for voice activity detection. Although different types of features are used in the literature to represent the video signal for voice activity detection, we remark on two main characteristics of speech, which are emphasized in these methods: the shape of the mouth and its dynamics. While open mouth is typically associated with speech intervals, the mouth can be completely closed during continuous speech. The second characteristic, which is the dynamics of the mouth movement, is used to avoid wrong detections in such scenarios by assuming rapid movement of the mouth during speech intervals.

Despite the large progress in the field of voice activity detection based on single channel processing (audio or video), only few studies focus on the incorporation of the audio and the video signals. In the study presented in [35] the audio and the video signals are mutually used for speech recognition. The audio and the video signals are represented using diffusion maps and are merged by concatenating the diffusion maps of each modality into a super-vector. Although the authors in [35] reported on promising results, the diffusion maps is limited in the representation of speech signals in the presence of non-speech interruptions and noise. In voice activity detection for example, the representation of transients may be similar to the representation of the clean signal (speech), which may lead to incorrect detections. In this study, we deal with this limitation by applying

diffusion maps in a domain of features which are designed to separate speech and non-speech frames. In particular, the kernel in (2.12), which represents the similarity between data points, is defined in the features domain, and the method reveals the geometric structure of the features, leading to a representation where speech and non-speech frames are separated into two different clusters. In addition, in Chapter 4 we show that the merging scheme proposed in [35, 37] is not optimal for voice activity detection in a noisy environment, and propose a merging scheme which provides improved performance.

In addition, in this chapter we reviewed the scattering transform which is exploited in Chapter 4 for voice activity detection. The scattering transform extends the representation of audio signals using MFCCs through a deep network, and allows for a better representation of transients using high orders of the scattering transform.

Chapter 3

Audio-Visual VAD Using Diffusion Maps

3.1 Introduction

In this Chapter, we present an algorithm for audio-visual voice activity detection. The inputs to the algorithm are audio and video signals recorded in a single microphone and a single video camera, respectively. The algorithm is based on a supervised learning procedure, and we consider a training data set which comprises speech signals contaminated by different types of noise and transients, and is labeled according to the presence and the absence of speech. The algorithm comprises three steps: first, high dimensional features are computed to represent the signals of each modality. The audio features are based on weighted MFCC [38] and are designed to separate the stationary from the non-stationary parts of the signal. The video features are based on motion vectors, which capture well both the shape of the mouth and its dynamics. Second, we adopt similar concepts to the spectral clustering algorithm presented in [16] and exploit a manifold learning method, diffusion maps [36], which is applied separately and similarly to the features computed from each modality. Diffusion maps provides a low dimensional representation of the signals which is suitable for merging data captured from different types of sensors [35]. In addition, it captures the intrinsic structure of the data and provides a good distance metric to separate speech and non-speech frames. Finally, a measure for voice activity is defined based on the diffusion mapping. This measure incorporates both a supervised

clustering procedure, which is based on a GMM, and an unsupervised procedure that exploits the variability of consecutive frames. The GMM is used to separate speech and non-speech clusters according to the labeled training data, and the unsupervised procedure separates the two clusters by assuming high variability between adjacent speech frames. The computed measures for voice activity from the two modalities are merged into a single bimodal measure, which is in turn used to estimate the speech presence indicator.

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

The remainder of the Chapter is organized as follows. In Section 3.2 we formulate the problem. The proposed algorithm is described in Section 3.3. Experimental results demonstrating the performance of the algorithm are presented in Section 3.4.

3.2 Problem Formulation

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

$$a[n] = a^s[n] + a^d[n] + a^t[n] \tag{3.1}$$

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

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

The video signal is assumed to comprise the region of the mouth, which is cropped out from a recorded front side video of a speaker. Note that the identification of the mouth region extends the scope of this study. Nevertheless, we briefly describe in section 3.4 the

procedure performed in our experiments as a preprocessing stage. Let $\mathcal{V} = \{\mathbf{v}_i\}_{i=1}^{N}$ be the video data set comprising N consecutive video frames $\mathbf{v}_i \in \mathbb{R}^{W \times H}$, where W and H are the width and the height of the frame, respectively.

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

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

$$\mathbb{1}_{\mathbf{s}}(i) = \left\{ \begin{array}{ll} 1, & i \in \mathcal{H}_1 \\ 0, & i \in \mathcal{H}_0 \end{array} \right\}. \tag{3.2}$$

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

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

3.3 Proposed Algorithm

An audio-visual voice activity detection algorithm consisting of three stages is proposed. First, audio and visual feature vectors are extracted. Then, a dimensionality reduction method, based on manifold learning [36], is applied to the extracted feature vectors. This provides a new low dimensional representation of the signals, which is in turn used for the estimation of the speech presence indicator.

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

learning) is denoted by $\{\hat{\mathbf{a}}_i, \hat{\mathbf{v}}_i\}$.

3.3.1 Features Extraction

3.3.1.1 Audio Features

The proposed audio features are based on spectral representation of speech using MFCCs and the STFT. These features were found to perform well for voice activity detection in challenging conditions, e.g., with a highly non-stationary noise [16].

Let $\mathbf{a}_j^{MFCC} \in \mathbb{R}^C$ be a column vector consisting of the MFCCs of frame $\mathbf{a}_j$, where C is the number of the coefficients. MFCCs are widely used in the field of speech recognition, since they successfully represent the spectrum of speech in a compact form using the perceptually meaningful Mel-frequency scale [38]. However, the MFCC representation of a speech frame may be similar to the representation of a non-speech frame comprising highly non-stationary noise. To improve the separation between the signal and the background noise, the MFCCs of each frame are weighted by a scalar which is based on noise estimation in the frame, such that a low value is assigned when only the background noise is present.

Traditionally, the input signal is assumed to contain only speech and stationary noise. Thus, speech and noise are separated assuming that stationary noise components in the STFT domain are slowly varying with respect to speech [11], [19]. However, transients vary faster than speech, and hence, they are mistakenly identified as (non-stationary) speech [15], [20]. Therefore, the weights which are based on the noise estimation method presented in [19], in this case, only separate speech and transients from the stationary (or quasi-stationary) noise. Next, we describe the computation of such weights and explain how to reduce the effect of transients on the frame representation.

Let $A(j,k)$ be the STFT representation of the audio signal $a[n]$, where k is the frequency bin, and j is the time frame index. Accordingly, the representation of (3.1) in the STFT domain is given by:

$$A(j,k) = A^s(j,k) + A^d(j,k) + A^t(j,k) \tag{3.3}$$

where $A^s(j,k)$, $A^d(j,k)$ and $A^t(j,k)$ are the STFTs of $a^s[n]$, $a^d[n]$ and $a^t[n]$, respec-

tively. The corresponding variances are given by $\lambda_s(j,k) = E\left[|A^s(j,k)|^2\right]$, $\lambda_d(j,k) = E\left[|A^d(j,k)|^2\right]$, and $\lambda_t(j,k) = E\left[|A^t(j,k)|^2\right]$, where $E[\cdot]$ denotes an expected value.

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

$$\Lambda_j(k) = \log\left(\frac{p_r\left(A(j,k);\mathcal{H}_{ns}\right)}{p_r\left(A(j,k);\mathcal{H}_s\right)}\right). \tag{3.4}$$

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

$$\xi(j,k) = \frac{\lambda_s(j,k) + \lambda_t(j,k)}{\lambda_d(j,k)}, \tag{3.5}$$

and is estimated according to [11], and let $\gamma(j,k)$ be the a posteriori NSSNR:

$$\gamma(j,k) = \frac{|A(j,k)|^2}{\lambda_d(j,k)}. \tag{3.6}$$

The estimation of both the a priori and the a posteriori NSSNRs is based on the spectral variance of the background noise, $\lambda_d(j,k)$, which is estimated using the Improved Minima Controlled Recursive Averaging (IMCRA) method [19].

Assuming that the non-stationary signal and noise have a complex uncorrelated Gaussian distribution in the STFT domain (using only speech and stationary noise model) it can be shown that the log likelihood ratio in (3.4) is given by [17], [5]:

$$\Lambda_j(k) = \left(\frac{\gamma(j,k)\cdot\xi(j,k)}{1+\xi(j,k)} - \log\left(1+\xi(j,k)\right)\right). \tag{3.7}$$

Let Λ_j be the arithmetic mean of the log likelihood ratio over all frequency bins of frame $\mathbf{a}_j$. To reduce the dynamical range of Λ_j, which is large, since, for example, when the background noise is absent, $p_r\left(A(j,k);\mathcal{H}_s\right) \to 0$ in (3.4), the weight of each frame $\mathbf{a}_j$ is given by normalizing Λ_j as:

$$w_\Lambda(j) = 1 - e^{-\frac{\Lambda_j}{\epsilon}}, \tag{3.8}$$

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

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

$$\tilde{\mathbf{a}}_i = \begin{pmatrix} w_\Lambda\left(i - J^A\right) \cdot \mathbf{a}_{i-J^A}^{MFCC} \\ w_\Lambda\left(i - J^A + 1\right) \cdot \mathbf{a}_{i-J^A+1}^{MFCC} \\ \vdots \\ w_\Lambda\left(i + J^A\right) \cdot \mathbf{a}_{i+J^A}^{MFCC} \end{pmatrix} \in \mathbb{R}^{(2J^A+1)\cdot C}. \tag{3.9}$$

It is worthwhile noting that Λ_j was previously used for voice activity detection in [5] and [6]. However, since Λ_j and $w_\Lambda(j)$ cannot exclusively indicate speech activity in the presence of transients, $w_\Lambda(j)$ is used in this study as a feature that separates speech and transients from background noise. Transient effects are attenuated by taking into account several consecutive time frames. This reduces the influence of transients on a frame representation since the typical duration of a transient is assumed to be of the order of a single time frame. Thus, for $J^A \geq 1$, the coordinates of $\tilde{\mathbf{a}}_i$ in the presence of speech are expected to be more consistent than in the presence of transients. In practice, we assign relatively small values to J^A, since large J^A induces a high dimension of features and requires a large number of training samples to construct the low dimensional model.

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

3.3.1.2 Visual Features

The proposed visual features are based on motion vectors which were previously exploited for voice activity detection in [26] and [27], and are calculated using Lucas-Kanade method [40], [41]. Let $\mathbf{v}_i(j,k)$ denote the (j,k)th pixel of frame v_i, and let $v_i(j,k)$ and $u_i(j,k)$ denote the horizontal and the vertical components of the motion vector (i.e., the velocity) of the corresponding pixel. We form a vector $\mathbf{v}_i^{MV} \in \mathbb{R}^{W\cdot H}$ by concatenating the absolute

values of the velocities of each pixel, which are given by $\sqrt{[v_i(j,k)]^2 + [u_i(j,k)]^2}$.

The video signal is characterized both by spatial information, i.e., the shape of the mouth, and by temporal information, i.e., the movement of the mouth. The shape of the mouth indicates on the presence of speech as the pronunciation of most of the phonemes is associated with open mouth [28]. However, the shape of the mouth cannot exclusively indicate on the presence of speech since, for example, the mouth can be completely closed in particular speech frames. Thus, temporal information may serve as a complement, i.e., the mouth movement may correctly indicate on the presence of speech. To capture both spatial and temporal information, motion vectors are calculated in a spatio-temporal neighborhood of each pixel in a frame.

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

$$\tilde{\mathbf{v}}_i = \left(\mathbf{v}^{MV}_{i-J^V}, \mathbf{v}^{MV}_{i-J^V+1}, \ldots, \mathbf{v}^{MV}_{i+J^V}\right)^T \in \mathbb{R}^{(2J^V+1)\cdot W\cdot H}. \tag{3.10}$$

Similarly to the parameter J^A in (3.9), the parameter J^V is set to a small value to confine the dimensions of the video features.

3.3.2 Diffusion Maps

Our main assumption is that the signals, which are captured by different types of sensors (microphone and camera), contain a common intrinsic low dimensional geometric structure, which is related to the speech activity. As a result, the high dimensional feature vectors are not spread across their entire space, but rather lie on a manifold of a significantly lower dimension.

In order to capture this low dimensional structure, we use diffusion maps [36], which is a manifold learning method, that provides a parameterization of the data on the manifold through the embedding of the high dimensional feature vectors into a low dimensional space. In this study we propose to apply diffusion maps in the domain of the features which are specifically designed to separate speech from non-speech frame. Therefore,

the diffusion maps reveals the manifold of the feature leading to a representation which naturally separates the speech and the non-speech cluster. In addition, diffusion maps is implemented by first constructing an empirical model of the manifold of the data using a training set, and then, the model of the manifold is extended to the test set in a frame by frame manner.

3.3.2.1 Construction of the empirical model using the training set

The method of diffusion maps is applied similarly and separately to the feature vectors of each modality. Let $\tilde{\mathbf{f}}_i$ be the feature vector (audio or visual) of the ith frame and let $\tilde{\mathcal{F}} = \left\{\tilde{\mathbf{f}}_i\right\}_{i=1}^{N}$ be the corresponding feature set. A pairwise similarity kernel function $k\left(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_j\right)$ between the ith frame and the jth frame is defined as:

$$k\left(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_j\right) = e^{-\frac{||\tilde{\mathbf{f}}_i - \tilde{\mathbf{f}}_j||^2}{\varepsilon}}.$$

Following this kernel, that defines similarity between two frames according to the L_2 distance in the features domain, we construct diffusion maps, described in equation (2.13)-(2.18) in Chapter 2. The diffusion maps of the ith training frame are given by:

$$\hat{\mathbf{f}}_i = \left(\mathbf{\Phi}_{i,1}, \mathbf{\Phi}_{i,2}, ..., \mathbf{\Phi}_{i,K}\right), \tag{3.11}$$

where $\mathbf{\Phi} \in \mathbb{R}^{N^{tr} \times K}$ is the matrix of the K largest eigenvalues and the corresponding eigenvectors of the transition probability matrix described in (2.17).

Then, the model of the manifold, which is constructed in a batch manner using the training set, is extended to new incoming frames. The extension is performed in a frame by frame manner similarly to the Nyström method [35], [42].

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

$$\phi'_k = \mathbf{w}^T \mathbf{\Phi}(:, k) \tag{3.12}$$

where $\mathbf{w}^T$ is the transpose of the weighting vector and $\mathbf{\Phi}(:, k)$ is the kth column of $\mathbf{\Phi}$. The extension may be seen as a weighted nearest neighbor interpolation, where $\mathbf{w}$ consists of

the interpolation weights. The ith entry of the weighting vector represents the similarity between the incoming test frame $\mathbf{f}_m$ and the ith training frame $\mathbf{f}_i$. Thus, the closer the extended frame is to a particular training frame in the features domain, the higher the weight of the diffusion maps entry of the training frame is in the extension. Traditionally, when the Nyström method is used to extend eigenvectors of a matrix, in our case $\mathbf{M}$, $\mathbf{w}(i)$ is given by $\mathbf{M}_{i,m}$. However, due to the normalization applied in (2.13)–(2.16), $\mathbf{M}_{i,m}$ can not be properly calculated in a frame-by-frame manner. Therefore, the "true" interpolation weights are approximated by a Gaussian kernel with the following correction [35]. Let $\mathbf{k} \in \mathbb{R}^{N^{tr}}$ be a vector whose ith entry is given by a Gaussian kernel:

$$\mathbf{k}_i = e^{-\left(\frac{||\tilde{\mathbf{f}}_i - \tilde{\mathbf{f}}_m||}{\sigma}\right)^2} \tag{3.13}$$

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

$$\mathbf{K}_{i,j} = e^{-\left(\frac{||\tilde{\mathbf{f}}_i - \tilde{\mathbf{f}}_j||}{\sigma}\right)^2} \tag{3.14}$$

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

$$\mathbf{w} = \mathbf{K}^{-1}\mathbf{k} \tag{3.15}$$

where the inverse matrix $\mathbf{K}^{-1}$ is used to correct the Gaussian weights. The weights are designed to provide a consistent extension of the diffusion maps. Namely, by substituting a training frame $\tilde{\mathbf{f}}_j$ instead of the test frame $\tilde{\mathbf{f}}_m$ into (3.14), the interpolation weight is degenerated to the Kronecker delta function, i.e., $\mathbf{w}_i = \delta_{i,j}$, and the extended value coincides with the true value, i.e., $\phi'_k = \mathbf{\Phi}(j,k)$.

Based on equation (3.12), the low dimensional representation of the test frame is given by:

$$\hat{\mathbf{f}}_m = (\phi'_1, \phi'_2, ..., \phi'_K)\,. \tag{3.16}$$

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

for real time applications.

3.3.3 Estimation of the speech presence indicator

3.3.3.1 Diffusion distance

Let $D(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_j)$ denote the diffusion distance between a pair of feature vectors $\tilde{\mathbf{f}}_i$ and $\tilde{\mathbf{f}}_j$, which is given by [37], [36]:

$$D^2(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_j) = \sum_{k=1}^{N^{tr}} \frac{(\mathbf{M}_{i,k} - \mathbf{M}_{j,k})^2}{\nu_0(k)} \tag{3.17}$$

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

$$\nu_0(j) = \frac{s\left(\tilde{\mathbf{f}}_j\right)}{\sum\limits_{\tilde{\mathbf{f}}_k \in \tilde{\mathcal{F}}^{tr}} s\left(\tilde{\mathbf{f}}_k\right)}. \tag{3.18}$$

The diffusion distance reflects the connectivity of the nodes (feature vectors) in the graph: Short distances are obtained for highly connected nodes due to high values of transition probabilities between the nodes [36]. The diffusion distance is known to be more robust to noise compared to the Euclidean distance, since it integrates information from many features, whereas the Euclidean distance takes into account only two individual features. In addition, the diffusion distance is unitless as it is calculated through transition probabilities. Therefore, it is suitable for merging data captured in different types of sensors.

When all the eigenvalues and the eigenvectors are used for the construction of diffusion maps in (2.18), i.e., $K = N^{tr}$, the L_2 distance in the diffusion maps domain equals the diffusion distance [36], [37]. Yet, even relatively small values of K (the dimension of the diffusion maps) provide an accurate approximation of the diffusion distance due to the fast spectrum decay:

$$D^2(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_j) \approx ||\hat{\mathbf{f}}_i - \hat{\mathbf{f}}_j||^2. \tag{3.19}$$

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

The new representation is particularly suitable for estimating the speech presence indicator both because of the specific choice of the feature vectors that characterize speech,

and because of the properties of the diffusion mapping. The later provides a low dimensional representation that captures the essence of the data and a good distance metric to compare embedded signal samples based on their intrinsic structure.

3.3.3.2 Unimodal estimation of speech presence indicator

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

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

$P^S(\mathbf{f}_i)$ is computed using a GMM procedure applied in the diffusion maps domain. Let $p_r(\cdot;\mathcal{H}_1)$ and $p_r(\cdot;\mathcal{H}_0)$ be the Gaussian mixture PDFs of the speech and the non-speech clusters, respectively, which are estimated using the expectation-maximization (EM) algorithm during the training stage [43]. Given a test frame $\mathbf{f}_i$, a bounded likelihood ratio between the conditional densities is calculated:

$$\Gamma_i = \min\left(\frac{p_r(\hat{\mathbf{f}}_i;\mathcal{H}_1)}{p_r(\hat{\mathbf{f}}_i;\mathcal{H}_0)}, \Gamma_{\max}\right), \tag{3.20}$$

where $\Gamma_{\max}$ is a constant value which is empirically set to confine the dynamical range of the likelihood ratio. The closer $\hat{\mathbf{f}}_i$ is to the speech training cluster, the higher Γ_i level is.

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

$$P^S(\mathbf{f}_i) = \frac{1}{(2L^S+1)\cdot\Gamma_{\max}} \sum_{r=-L^S}^{L^S} \Gamma_{i+r}. \tag{3.21}$$

Recall that in (3.9) and in (3.10), the use of the temporal neighborhood to characterize speech was limited to keep reasonable values of the dimensions of the features. In (3.21), the temporal neighborhood is exploited without these limitations, and Γ_i is averaged

over $(2L^S + 1)$ consecutive frames to smooth the measure of voice activity and improve the estimation of speech presence indicator. For example in non-speech intervals, short term interruptions such as transients may provide instantaneous high values of Γ_i, yet, smoothing the measure over the temporal neighborhood provides correct low levels of voice activity.

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

$$P^{US}(\mathbf{f}_i) = \frac{\min\left(\sum_{l=1}^{L^{US}} D(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_{i-l}), \sum_{l=1}^{L^{US}} D(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_{i+l})\right)}{L^{US} \cdot D_{\max}}, \tag{3.22}$$

where $D_{\max}$ is the maximal diffusion distance between a pair of frames in the training set, and $L^{US} \cdot D_{\max}$ is a normalization factor, which keeps P^{US} values (given a large enough training set) in the range of $[0,\ 1]$. In speech absence periods, audio frames tend to be similar to their adjacent frames as background noise varies slower than speech. A similar property is observed in video frames, as a slower mouth movement is assumed in speech absence periods compared to periods when speech is present. Accordingly, P^{US} is expected to provide lower levels of voice activity when speech is absent compared to when it is present, for both modalities. According to (3.22) the variability of the frames is measured in two non-overlapping windows: a causal window and an anti-causal window, both of size L^{US}. The min function is used to reduce false detection at the beginning and at the end of speech intervals. For example, correct low values of activity are received right after the end of a speech interval due to low values of variability in the anti-causal window despite high levels of variability in the causal window. Speech presence estimation based on P^{US} is viewed as an unsupervised procedure since the training data is used only for the construction of diffusion maps without its labeling.

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

$$P(\mathbf{f}_i) = \min\left(P^S(\mathbf{f}_i), P^{US}(\mathbf{f}_i)\right). \tag{3.23}$$

The min function is used to reduce false high activity levels of P^S or P^{US}. For exam-

ple, highly non-stationary noise such as "babble" noise induces high variability between consecutive time frames in non-speech periods, which may be resulted in false high levels of P^{US}. Yet, successful distinction of this type of noise in the diffusion domain provides accurate P^S. Hence, in such cases, the min function may improve the estimation of the speech presence indicator.

3.3.3.3 Bimodal estimation of speech presence indicator

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

$$P^B(\mathbf{a}_i, \mathbf{v}_i) = \alpha P(\mathbf{a}_i) + (1 - \alpha) P(\mathbf{v}_i) \tag{3.24}$$

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

In this work, for simplicity we set $\alpha = 0.5$. Combining the modalities this way was derived in [44] through a Bayesian model under restrictive assumptions that the modalities are statistically independent and that a posteriori probability of each modality remains close to the priors. Nevertheless, it was empirically found to outperform other fixed functions for combining the two modalities (such as a product, minimum, maximum, and median) due to better resistance to estimation errors. This simple combination empirically showed good results as illustrated in Section 3.4. Adaptive setting of α will be addressed in a future work.

A different approach for combining the modalities can be achieved by concatenating the diffusion maps of each modality into a single super-vector [37], [35]. As a result, the speech presence indicator can be estimated in a unified diffusion maps domain, and $P(\cdot)$ in (3.23) represents a bimodal measure for speech presence. We will show in Section 3.4

that this approach provides inferior results with respect to the proposed algorithm.

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

$$\hat{\mathbb{1}}_{\mathbf{s}}(i) = \left\{ \begin{array}{ccc} 1 & ; & P^B(\mathbf{a}_i, \mathbf{v}_i) > \tau \\ 0 & ; & otherwise \end{array} \right\}. \tag{3.25}$$

Future frames which are used in (3.9), (3.10), (3.21) and (3.22) induce a lag in online processing. The number of lagged frames, N^{lag}, is given by:

$$N^{lag} = \max\left(J^A, J^V\right) + \max(L^S, L^{US}) \tag{3.26}$$

The effect of the lag on real time processing is discussed in Section 3.4.

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

3.4 Experimental Results

3.4.1 Experimental setup

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

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

Although cropping the bounding box of the mouth extends the scope of this work, we

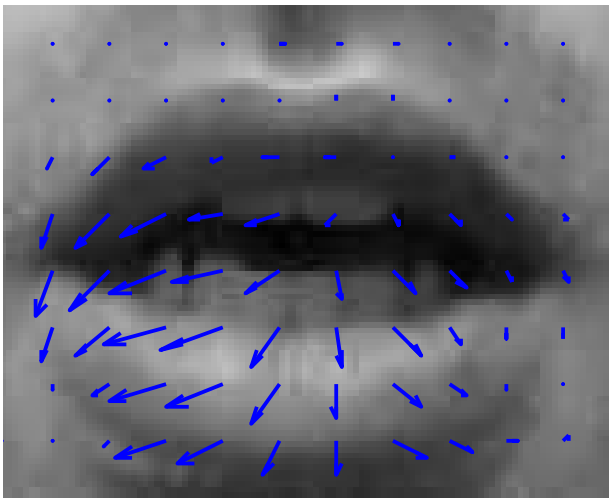


Figure 3.1: Motion vectors in a speech interval and the corresponding video frame.

shortly explain the procedure performed in our experiments as a preprocessing stage. The cropping is based on nostrils tracking, which are manually marked in the first frame, and are then searched in the following frames in a small area around their previous location. The search is performed under the assumption that the pixels, where the nostrils are located, have lower intensity values relatively to skin and lips pixels due to shading. Such a nostrils tracking method was previously explored in [45].

The bounding box of the mouth is downsampled by a factor of 10 to reduce the computational load in the calculation of the motion vectors, and W and H are set to 11 and 9 (all the used parameters values are presented in Table 3.1). An example of a speech frame image and an illustration of the corresponding motion vectors are presented in Fig. 3.1, demonstrating that motion vectors capture the shape of the mouth and its movement with respect to the previous frame.

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

The training data set is created by collecting 30 s long data sequences of 6 speakers (the total training data set is 180 s, 4542 frames). We empirically find that a 180 [sec] long signal both contains sufficient amount of training data and the eigenvalue decomposition

Table 3.1: Algorithm Parameters

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

can be efficiently applied to $\mathbf{M}$ (using Intel® Core™ i5-2500 CPU and 4 GB RAM). To make the calculation more efficient, $\mathbf{M}$ may be processed in blocks similarly to [16].

The algorithm is trained for challenging acoustic environments: various background noise types, which include white Gaussian noise, musical noise, and babble noise, and various transient interferences, such as metronome, keyboard typing, and hammering, taken from [46], are added to the training audio signal of each speaker. The algorithm is trained for 0 and 5 dB SNR values, and the transients are normalized to have maximal amplitude twice larger than the maximal amplitude of speech. The training data of each speaker contains all possible combinations of background noise and transients. This training setup extends the setup in [16], where only a single background noise and a single transient type were used for the training in each experiment.

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

3.4.2 Qualitative evaluation

The 20 largest eigenvalues of the audio and video sets, $\{\lambda_k\}$ in (2.17), are plotted in a decreasing order in Fig. 3.2, demonstrating fast decay of the spectrum. A spectral gap can be seen between the fourth and fifth eigenvalues for the audio signal. This gap,

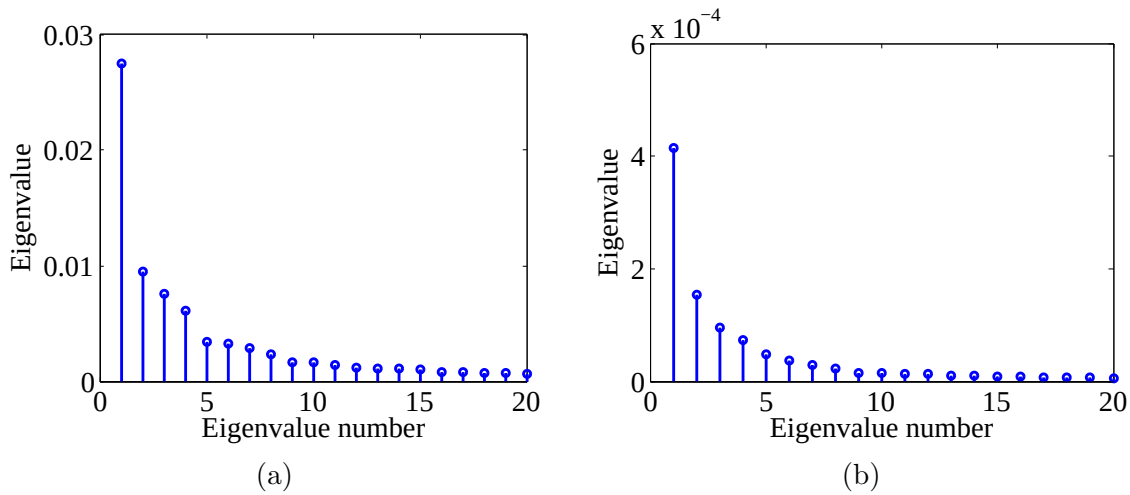


Figure 3.2: 20 largest eigenvalues of transition probability matrix, (a) audio, (b) video.

heuristically implies that the intrinsic dimension of the signal is 4 [47]. Consequently, we set the dimension of diffusion maps to $K = 4$ in (2.18) for each modality. The choice of 8 parameters (4 for each modality) for representing a frame yields a significant dimensionality reduction of the data, and hence, allows for low computational complexity of the estimation procedure. In addition, these parameters capture just the essence of the data without noise and other nuisance factors, thereby allowing accurate detection of voice activity. The difference in the visual data representation from most of the previous studies, such as [28] and [29], is that in this work the parameters representing the mouth movement are obtained implicitly in a data-driven manner and are not defined in advance.

The measures of voice activity P^S in (3.21) and P^{US} in (3.22) are calculated using $L^S = L^{US} = 9$ past and future frames. This configuration induces a lag of $N^{lag} = 10$ frames in (3.26), which is ~ 400 ms for ~ 25 fps frame rate. We empirically found that lower values of lag may be set at the expense of a small degradation of the performance.

An example of the obtained voice activity detection is shown in Fig. 3.3. The input signal (black solid line) in this example is contaminated with a 10 dB babble noise and keyboard typing transients. We observe an accurate speech presence indicator estimation when compared to the marked ground truth. We note that in this example, the threshold τ in (3.25) is empirically chosen to provide best estimation results. Although it is not in the scope of this work, in practice, the threshold may be set using the training data by evaluating the performance of the algorithm using a validation set.

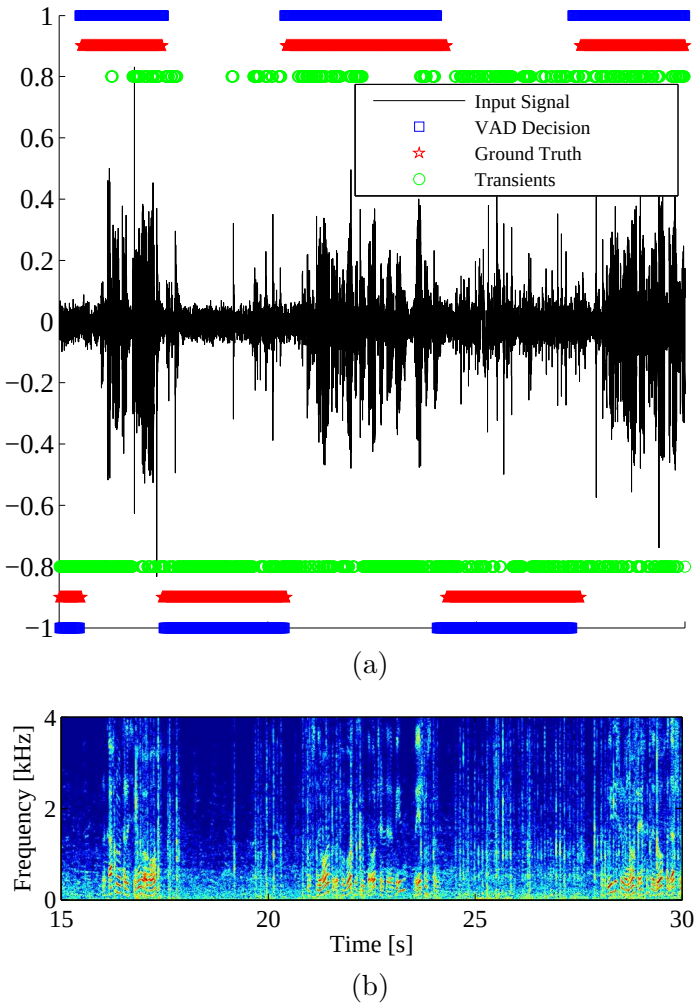


Figure 3.3: Qualitative assessment of AVVAD, with babble noise with 10 dB SNR and keyboard typing transient interferences. (a) Input signal- black solid line, ground truth - red stars graph, speech presence estimation using the proposed method - blue squars graph, the locations of the transients - green rings graph. (b) A spectrogram of the input signal.

3.4.3 Performance evaluation measure

For voice activity detection, the ground truth may be application dependent. For example, in speech recognition applications, isolated phonemes may be useful, and hence, the voice detection should ideally have a fine resolution in the order of few tens of milliseconds. On the other hand, in videoconferencing systems, where a central processing unit switches between cameras according to a dominant speaker [3], frequent switching between speakers should be avoided, and hence, coarser voice activity detection is required.

In addition, the ground truth depends on the modality (audio or video). Speech onsets, for example, may be accompanied with air aspiration, which is helpful for visual speech

perception, and therefore, is considered as speech for the video signal but not for the audio signal. Another example is voiced phonemes. While the ends of voice phonemes are audible, they are not visual due to the lack of a mouth movement.

Therefore, we extend the definition of the speech indicator in (3.2). Similarly to the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, let $\mathcal{H}_0^a$ and $\mathcal{H}_1^a$ be the hypotheses for speech absence and speech presence in the audio signal, respectively. Accordingly, let $\mathbb{1}_{\mathbf{s}}(\mathbf{a}_i)$ be an audio speech indicator, given by:

$$\mathbb{1}_{\mathbf{s}}(\mathbf{a}_i) = \left\{ \begin{array}{ccc} 1 & ; & \mathbf{a}_i \in \mathcal{H}_1^a \\ 0 & ; & \mathbf{a}_i \in \mathcal{H}_0^a \end{array} \right\}. \tag{3.27}$$

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

$$\mathbb{1}_{\mathbf{s}}\left(\mathbf{v}_i\right) = \left\{ \begin{array}{ccc} 1 & ; & \mathbf{v}_i \in \mathcal{H}_1^v \\ 0 & ; & \mathbf{v}_i \in \mathcal{H}_0^v \end{array} \right\}. \tag{3.28}$$

The video speech indicator is manually marked as speech present when mouth moves during speech intervals (a natural mouth movement when speech is absent is neglected). The unified speech indicator $\mathbb{1}_{\mathbf{s}}(i)$ which is defined in (3.2) is given by:

$$\mathbb{1}_{\mathbf{s}}(i) = \mathbb{1}_{\mathbf{s}}\left(\mathbf{a}_i\right) \vee \mathbb{1}_{\mathbf{s}}\left(\mathbf{v}_i\right) \tag{3.29}$$

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

The quantitative performance is evaluated in three experiments. In the first and second experiments, unimodal versions of the proposed algorithm are compared to the state of the art and recently presented VADs. To evaluate the performance of the unimodal versions, $P^B(\mathbf{a}_i, \mathbf{v}_i)$ in (3.24) is replaced with the single modality activity measure P defined in (3.23): $P(\mathbf{a}_i)$ for audio and $P(\mathbf{v}_i)$ for video. For these experiments, the ground truth is given by (3.27) and (3.28) for evaluation based on only the audio signal and only the video signal, respectively. In the third experiment, the proposed AVVAD is compared to

the single modality versions using the audio-visual ground truth given by (3.29).

3.4.4 VAD evaluation

The performance of the proposed algorithm based on the audio signal is compared to the methods presented in [5] and [6]. Similarly to the proposed algorithm, in the algorithm presented in [6], the likelihood ratio is calculated in past and future frames. This allows for activity level smoothing and is more adequate for the ground truth setting in this work. However, in the VAD presented in [5], the likelihood ratio is calculated for a single frame, which makes the detection less adequate to this application. To make a fair comparison, we smooth the VAD in [5] with a median filter of length 11, which significantly improves its performance.

The proposed algorithm based on the video is compared to the method presented in [28]. Unlike the proposed algorithm, the VAD presented in [28] is not designed to perform in real time, as the parameters of the noise statistics in [28] are estimated in a batch manner. For simplicity of the implementation, these parameters are estimated using the ground truth of the test data set. We remark that estimation using the training set was also performed as suggested in [27]. Although it allows real time processing, our experiments show significant degradation of the performance of the competing algorithm, and hence, these results are not presented in the figures. In addition, we also evaluated the algorithm presented in [29]. We found that the procedure that separates the lips from the skin performed poorly on some of the videos in our data set, probably due to the lightening conditions. As a result, the overall performance of the algorithm was not comparable to the other two algorithms, and hence, is not presented in the figures.

Figures 3.4–3.8 present the obtained Receiver Operating Characteristic (ROC) curves, which are generated by spanning the threshold over all possible activity values. The maximal performance of each method, in terms of the number of correct decissions, is given in percents in the legend box of each figure and is obtained using the threshold which provides the best results.

In Figs. 3.4 and 3.5, we present the results of the evaluation of the algorithms based on the audio signal, where the curves marked by “Sohn” and “Ramirez” relate to the methods presented in [5] and [6], respectively. In Fig. 3.4, the algorithm is tested for babble noise

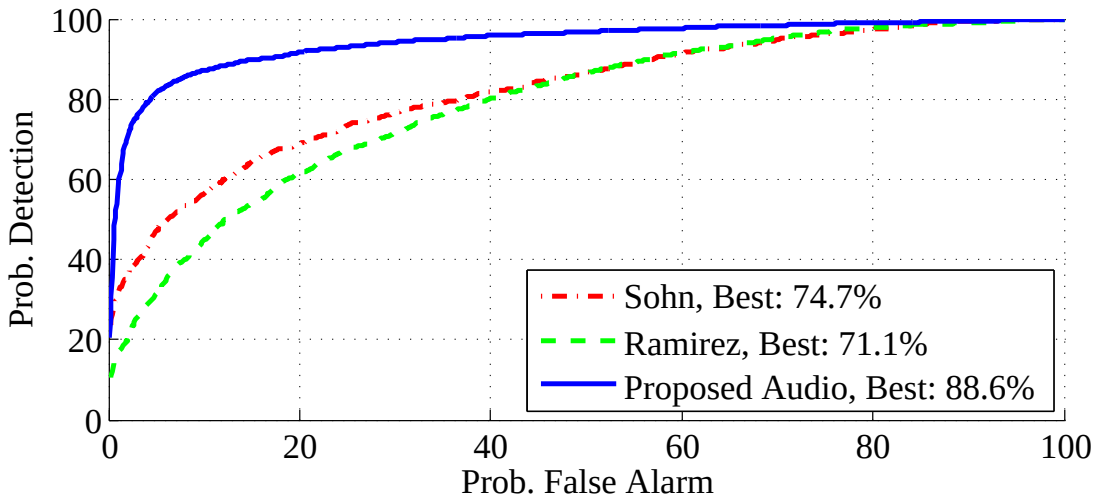


Figure 3.4: Audio algorithms. Probability of detection vs probability of false alarm. Test for babble noise with 5 dB SNR and keyboard typing transient interferences.

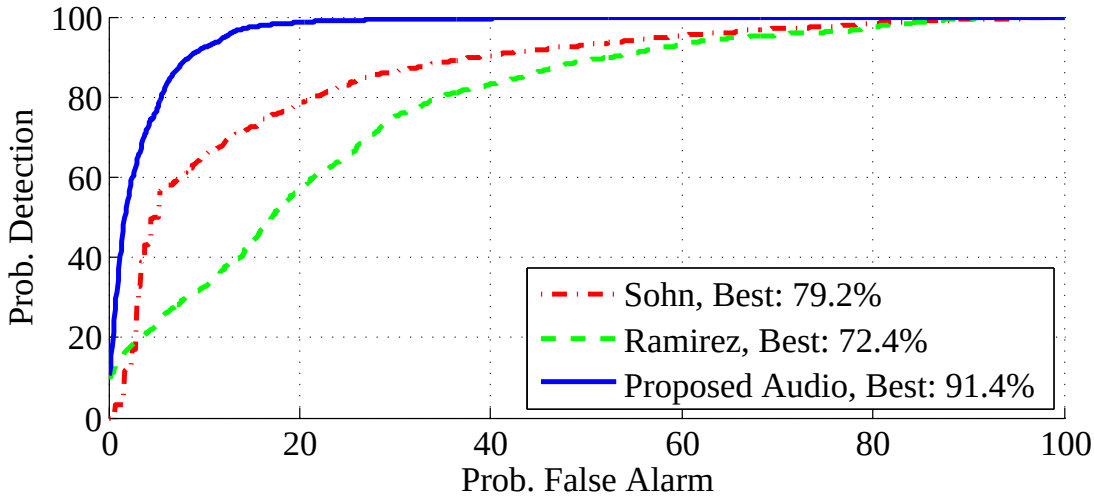


Figure 3.5: Audio algorithms. Probability of detection vs probability of false alarm. Test for musical noise with 0 dB SNR and scissors transient interferences

with 5 dB SNR and keyboard typing transients and in Fig. 3.5 for musical noise with 0 dB SNR and scissors transients. It can be seen in both figures that the proposed algorithm outperforms the competing audio VADs for all possible values of false alarm rates. We emphasize that the presented results of the proposed algorithm are achieved with a single training set consisting of all types of background noises and transients. Namely, the type of the tested background noise and transient are not known in advance.

The results of the algorithms based on the video signal are presented in Fig. 3.6, where the curve marked by "Siatras" relates to the method presented in [28]. It can be seen that the proposed algorithm outperforms the VAD in [28] for all possible values of false

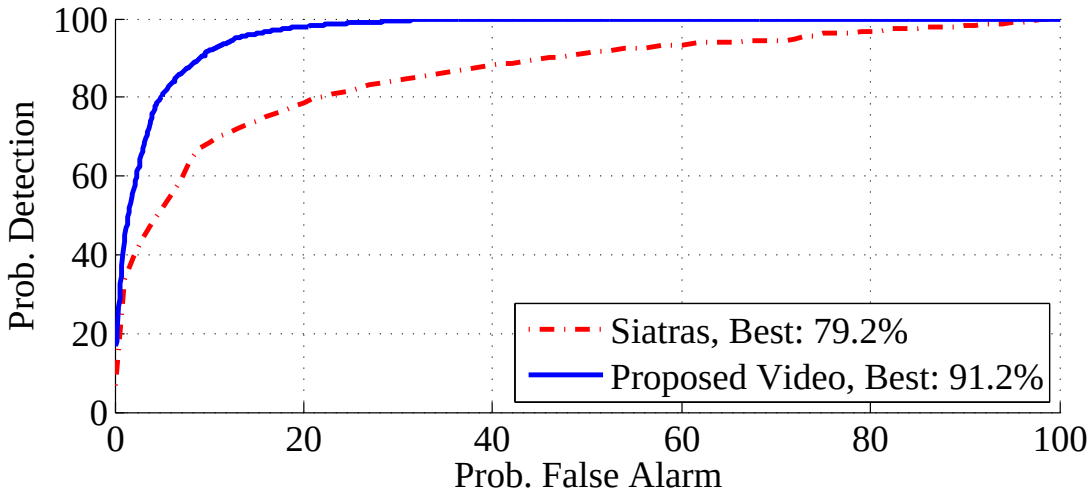


Figure 3.6: Video algorithms. Probability of detection vs probability of false alarm.

alarm rates. In addition, a high slope of the ROC curve of the proposed VAD is observed for low false alarm rates, thereby providing fast convergence to high detection rates.

In the third experiment, we evaluate the performance of the proposed AVVAD and compare it to the performance of the algorithm using single modalities. In addition, we evaluate the performance of an audio-visual merging approach which was exploited in [35] for speech recognition. In this approach, the diffusion maps of each modality are concatenated into a single super-vector, and is marked by "SV AV" in the plots.

The performances of the algorithms are presented in Figs. 3.7 and 3.8. In Fig. 3.7, the algorithms are tested for babble noise with 15 dB SNR and scissors transients, and in Fig. 3.8 the algorithms are tested for white Gaussian noise with 10 dB SNR and metronome transients. It can be seen that the merging scheme presented in [35] does not benefit from the combination of the modalities, and in most cases provides inferior performance with respect to the unimodal versions. This merging scheme is less adequate for merging data in a noisy environment since interruptions from each one of the modalities may wrongly lead to high levels of the bimodal voice activity measure. An example of incorrect classification is mouth movement during non-speech intervals. In this case, high diffusion distance from the non-speech cluster may be achieved due to the movement of the mouth, incorrectly providing high value of the voice activity measure. In contrast, the proposed bimodal algorithm performs better than each of the single modality versions. Moreover, it can be seen in both figures that the audio and the video versions may complement each other.

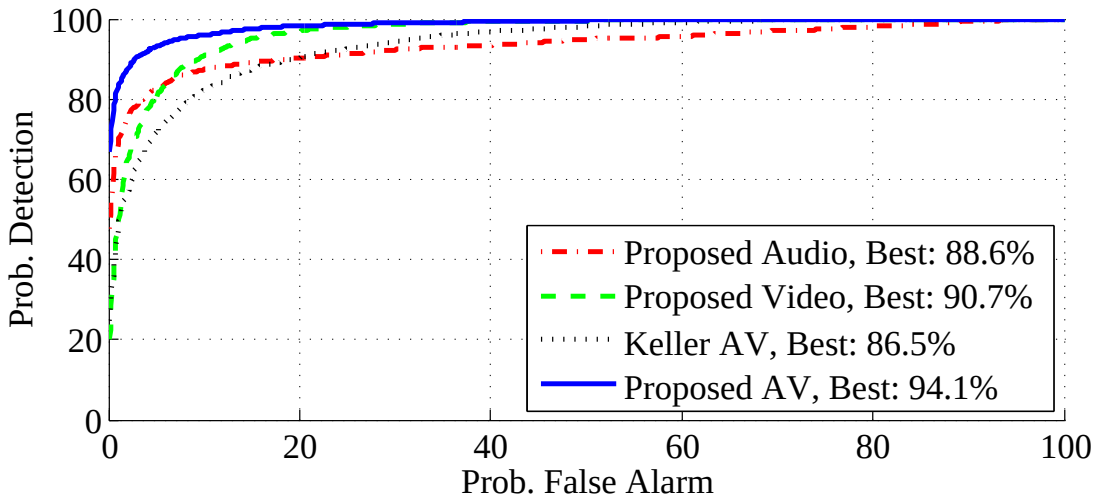


Figure 3.7: Audio, video and audio-visual algorithms. Probability of detection vs probability of false alarm. Test for babble noise with 15 dB SNR and scissors transient interferences.

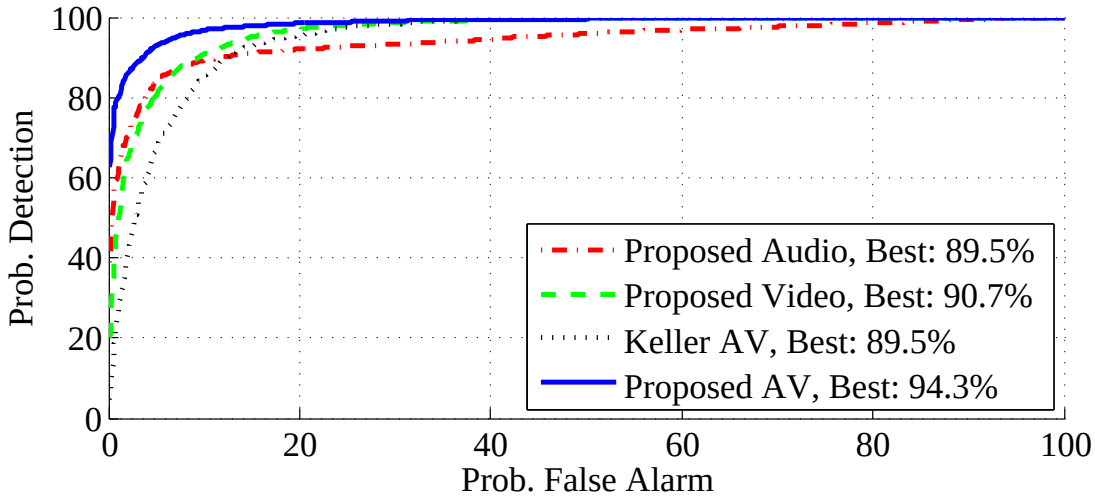


Figure 3.8: Audio, video and audio-visual algorithms. Probability of detection vs probability of false alarm. Test for white Gaussian noise with 10 dB SNR and metronome transient interferences.

While the audio version better performs for low values of false alarm, the video version is better for the high values. The proposed bimodal algorithm embodies the advantages of each of the modalities, and provides best performance for each false alarm value.

3.5 Conclusions

We have presented an algorithm for audio-visual voice activity detection. The algorithm is based on a low dimension representation of the audio and the video signals which is

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

Algorithm 3.1 Audio-visual voice activity detection

procedure Training
Input: training data- $\{\mathbf{a}_i, \mathbf{v}_i\}_{i=1}^{N^{tr}}$
Output: diffusion maps- $\{\hat{\mathbf{a}}_i, \hat{\mathbf{v}}_i\}_{i=1}^{N^{tr}}$; estimate of the PDFs, $p_r(\cdot;\mathcal{H}_0)$ and $p_r(\cdot;\mathcal{H}_1)$, for the GMM
for $j = 1 : N^{tr}$
  Calculate the feature vectors $\tilde{\mathbf{a}}_j$, $\tilde{\mathbf{v}}_j$
end for
Do for each modality separately:
Calculate the transition probability matrix $\mathbf{M}$ using (2.12)-(2.16)
Apply eigenvalue decomposition on $\mathbf{M}$ and obtain the eigenvalues $\{\lambda_k\}$ and the eigenvectors $\{\boldsymbol{\phi}_k\}$
Build diffusion maps $\hat{\mathbf{f}}_j, j \in [1, 2, ..., N^{tr}]$ ($\hat{\mathbf{a}}_j$ for audio and $\hat{\mathbf{v}}_j$ for video) using (2.17) and (2.18)
Train the GMM using the labeling and estimate $p_r(\cdot;\mathcal{H}_0)$ and $p_r(\cdot;\mathcal{H}_1)$
end procedure

procedure Test
Input: test data- $\{\mathbf{a}_i, \mathbf{v}_i\}_{i=1}^{N^{te}}$
Output: speech presence indicator estimate- $\{\hat{\mathbb{1}}_{\mathbf{s}}(i)\}_{i=1}^{N^{te}}$
* Get a new frame $\mathbf{a}_i$, $\mathbf{v}_i$
Calculate the feature vectors $\tilde{\mathbf{a}}_i$, $\tilde{\mathbf{v}}_i$
Extend the diffusion maps $\hat{\mathbf{a}}_i$, $\hat{\mathbf{v}}_i$ using (3.12) and (3.16)
Do for each modality separately:
Calculate the first voice activity measure $P^S(\mathbf{f}_i)$ using the PDFs of the GMM according to (3.20) and (3.21)
Calculate the second voice activity measure $P^{US}(\mathbf{f}_i)$ according to (3.22)
Integrate the measures:
$P(\mathbf{f}_i) = \min\left(P^S(\mathbf{f}_i), P^{US}(\mathbf{f}_i)\right)$
Merge the modalities:
$P^B(\mathbf{a}_i, \mathbf{v}_i) = \alpha P(\mathbf{a}_i) + (1-\alpha)P(\mathbf{v}_i)$
if $P^B(\mathbf{a}_i, \mathbf{v}_i) > \tau$
Decide $\mathcal{H}_1$
else
Decide $\mathcal{H}_0$
end if
Go back to *
end procedure

Chapter 4

Audio VAD Using The Scattering Transform

4.1 Introduction

In this chapter, we present a supervised learning algorithm for voice activity detection in the presence of highly non-stationary noise and transients. To train the algorithm, we consider a labeled training data set of speech signals contaminated with noise and transients. The algorithm is based on the representation of the noisy signal by features which are based on the scattering transform [1, 39, 48], include noise estimation, and are specifically designed to separate speech and non-speech frames. For voice activity detection, we propose a continuous measure which is constructed in the features domain, and rely on the SVM classifier. The algorithm is evaluated for different types of background noises and transients, and experimental results demonstrate improved voice activity detection compared to state-of-the-art methods. We remark that following [1], we present this chapter in the continuous domain.

The remainder of the Chapter is organized as follows. In Section 4.2 we formulate the problem. The proposed algorithm is described in Section 4.3. Experimental results demonstrating the performance of the algorithm are presented in Section 4.4.

4.2 Problem Formulation

Let $y(\tau)$ denote a speech signal contaminated with additive background noise and an additive transient interference, given by:

$$y(\tau) = x(\tau) + d(\tau) + z(\tau) \tag{4.1}$$

where $x(\tau)$,$d(\tau)$ and $z(\tau)$ are speech, the background noise and the transient interference, respectively. The signal is processed in overlapping time frames of length T, such that time frame t is denoted by y_t and is given by $y(\tau)$; $\tau \in [t - T/2\,,\, t + T/2]$. Let $\mathbb{1}_{\mathbf{s}}(t)$ denote a speech indicator of frame t, given by:

$$\mathbb{1}_{\mathbf{s}}(t) = \left\{ \begin{array}{ccc} 1 & ; & t \in \mathcal{H}_1 \\ 0 & ; & t \in \mathcal{H}_0 \end{array} \right\} \tag{4.2}$$

where $\mathcal{H}_1$ and $\mathcal{H}_0$ are two hypotheses denoting speech presence and absence, respectively. The goal in this study is to estimate the speech indicator in (4.2) for each frame. The algorithm is based on a supervised learning procedure, and we consider a training data set which consists of speech signals contaminated with noise and transients, and is labeled according to the speech absence and presence hypotheses.

4.3 Proposed Algorithm

4.3.1 The Features

The proposed features are based on the scattering transform which is a cascade of wavelet convolutions and modulus operators [1,48]. Let a wavelet $\psi(\tau)$ be a band pass filter with a central frequency normalized to 1, and let $\{\psi_\lambda(\tau)\}_\lambda$ be a wavelet filter bank, which is constructed by dilating the wavelet:

$$\psi_\lambda(\tau) = \lambda\psi(\lambda\tau), \tag{4.3}$$

where $\lambda = 2^{j/Q}$, $\forall j \in \mathbb{Z}$ and Q is the number of wavelets per octave. The bandwidth of the wavelet $\psi(\tau)$ is of the order of $1/Q$, and as a result, the filter bank is composed of band pass filters which are centered in the frequency domain in λ and have a frequency bandwidth λ/Q, i.e., they are logarithmically spaced in the frequency domain. The first and the second orders of the scattering transform are denoted by, $S_1(\tau, \lambda_1)$ and $S_2(\tau, \lambda_1, \lambda_2)$, respectively, and are given by:

$$S_1(\tau, \lambda_1) = |y * \psi_{\lambda_1}| * \phi(\tau), \tag{4.4}$$

and:

$$S_2(\tau, \lambda_1, \lambda_2) = ||y * \psi_{\lambda_1}| * \psi_{\lambda_2}| * \phi(\tau), \tag{4.5}$$

where $\phi(\tau)$ is a low pass filter with a frequency bandwidth $2\pi/T$. A scattering vector of frame t is denoted by $\mathbf{y}_t^S$ and is given by concatenating the first and the second order of the scattering transform calculated at time t for each filter.

The scattering transform is invariant to time shifts and is stable to time-warping due to the logarithmically spaced filter bank [48], making it useful for classification (see more details in [1, 48]). These properties are also held for the Mel-Frequency Spectral Coefficients (MFSC) obtained by averaging the signal in the STFT domain with Mel-scale filters which are also logarithmically spaced in the frequency domain for high frequencies [1]. In addition, it is shown in [1] that the MFSCs are similar to the coefficients of the first order of the scattering transform. The MFCCs are given by applying a cosine transform on the log of the MFSCs, they are widely used in speech recognition [38], and were recently exploited for voice activity detection in [16]. However, the averaging in the frequency domain in the construction of the MFSCs and MFCCs removes information over small time scales. Similarly, the convolution with the low pass filter in (4.4) causes loss of information. In particular, the representation of transients which are usually short in time, may be similar to the representation of speech, and may lead to false voice activity detection in the presence of transients. The second order of the scattering transform recovers the lost information using a new set of wavelet filters and the modulus operator [1,48]. Therefore, representation of signals using the first and the second orders of the scattering transform extends the MFCC representation and better separates between

speech and transients.

Yet, non-speech frames which are contaminated with background noise may be similar to speech frames. In order to improve the separation between speech and noise frames, the scattering vector $\mathbf{y}_t^S$ is weighted with a scalar which incorporates noise estimation in the STFT domain [16]. Let $Y(t,\omega)$ be the STFT of $y(\tau)$, and let $p_r(Y(t,\omega);\mathcal{H}_0)$ and $p_r(Y(t,\omega);\mathcal{H}_1)$ be PDF of the noisy signal conditioned on the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. The log of the likelihood ratio between the conditional PDFs is given by:

$$\Lambda_t(\omega) = \log\left(\frac{p_r(Y(t,\omega);\mathcal{H}_1)}{p_r(Y(t,\omega);\mathcal{H}_0)}\right). \tag{4.6}$$

Λ_t is a scalar obtained by averaging the log of the likelihood ratio in (4.6) over the frequency scale, and is used to weight the features. The weight of frame t is denoted by w_t and is given by:

$$w_t = 1 - e^{-\frac{\Lambda_t}{\epsilon}}$$

where ϵ is a normalization parameter. Accordingly, the feature vector of frame t is given by:

$$\mathbf{y}_t = w_t \mathbf{y}_t^S. \tag{4.7}$$

In frames which contain only background noise, Λ_t receives low values since $p_r(Y(t,\omega);\mathcal{H}_1) \to 0$ in (4.6), and ϵ is set such that w_t receives values close to 0. In speech frames, Λ_t in (4.6) receives high values, and w_t receives values close to 1. Λ_t is estimated according to [5] and incorporates noise estimation procedure which is based on the assumption that noise is (quasi) stationary and is slowly varying with respect to speech [17, 19]. Since transients are highly non-stationary signals and are varying faster than speech, high values of Λ_t are obtained also in presence of transients. As a result, w_t receives values close to 1 both in the presence of speech and transients, and is used in this work to separate noise from the non-stationary part of the signal, i.e., speech and transients. Therefore, the proposed features allows for the separation of speech from noise using the weighting scalar and from transients, using the second order of the scattering transform.

4.3.2 Voice Activity Detection

We base the estimation scheme on the SVM procedure. Originally, this procedure provides a binary classification of feature vectors according to their position with respect to a hyperplane, which is optimized using the labeled training data to maximize inter class separation. In this study, we propose a continuous measure for voice activity which is based on the distance of the tested features to the hyperplane such that the classification is given by comparing the measure to a threshold. The advantage of a continuous measure over a binary classification is that the threshold value, which controls the tradeoff between false alarm and correct detection rates, may be adjusted to a specific application. Let $\mathbf{n} \in \mathbb{R}^{\mathbb{K}}$ be the normal vector (not necessarily normalized) to the hyperplane, and let b be a parameter such that $b/||\mathbf{n}||$ is the offset of the hyperplane from the origin, where $||\cdot||$ is the L_2 norm. The distance of a tested feature vector $\mathbf{y}_t$ from the hyperplane, denoted by L_t, is given by:

$$L_t = \frac{< \mathbf{y}_t, \mathbf{n} > + b}{||\mathbf{n}||}.$$

Note that for simplicity, we relate to a linear SVM, while the extension to a kernel SVM is straightforward. In a binary classification, $\mathbf{y}_t$ is classified according to the sign of L_t such that $\mathbf{y}_t$ is considered as a speech frame if (say) $L_t > 0$ and as a non-speech frame otherwise. In this work we propose a continuous measure for voice activity which exploits the dynamical range of L_t rather than its sign, and in particular, we assume that large values of L_t indicate on high probability of voice activity in frame t. To define the voice activity measure, we first reduce the dynamical range of L_t by applying a soft threshold. The distance with a reduced dynamical range is denoted by $\acute{L}_t$ and is given by:

$$\acute{L}_t = \left\{ \begin{array}{ccc} L_{\min} & ; & L_t < L_{\min} \\ L_t & ; & L_{\min} < L_t < L_{\max} \\ L_{\max} & ; & L_t > L_{\max} \end{array} \right\}, \tag{4.8}$$

where $L_{\min}$ and $L_{\max}$ are constant distances from the hyperplane such that beyond them speech is assumed to be absent and present, respectively. $L_{\min}$ and $L_{\max}$ are empirically set to be half of the maximal negative and positive distances from the hyperplane in the training set, respectively. Then, $\acute{L}_t$ is normalized to provide values in the range of $0 \div 1$,

and the normalized distance, denoted by $\tilde{L}_t$, is given by:

$$\tilde{L}_t = \frac{\acute{L}_t - L^{\min}}{L^{\max} - L^{\min}} \tag{4.9}$$

The voice activity measure, denoted by P_t, is given by averaging $\tilde{L}_t$ over $2J+1$ temporally neighboring frames:

$$P_t = \frac{1}{2J+1} \sum_{j=t-JT}^{t+JT} \tilde{L}_j, \tag{4.10}$$

where J is a non-negative parameter that defines the temporal neighborhood. The value of P_t is in the range of $0 \div 1$, and the higher P_t the higher the probability for speech presence in frame t. By taking into account several consecutive frames in (4.10), the effect of transients on the voice activity measure is attenuated since their length is assumed to be of the order of a single frame. The speech presence indicator defined in (4.2) is estimated by comparing the speech presence measure P_t to a threshold α such that the estimated indicator, denoted by $\hat{\mathbb{1}}_{\mathbf{s}}(t)$, is given by:

$$\hat{\mathbb{1}}_{\mathbf{s}}(t) = \left\{ \begin{array}{ccc} 1 & ; & P_t > \alpha \\ 0 & ; & \text{otherwise} \end{array} \right\}. \tag{4.11}$$

4.4 Experimental Results

In this section we evaluate the performance of the proposed algorithm and compare it to the methods presented in [5], [6] and [16], which are called "Sohn", "Ramirez" and "Mousazadeh" in the plots, respectively. In addtion, we compare the proposed algorithm to the audio version of the VAD presented in Chapter 3, which is called "Aud DM" in the plots. The algorithm is evaluated for different types of background noises, including white Gaussian noise, colored Gaussian noise and babble noise, and different types of transients e.g., door knocks and keyboard taps. The SNR is defined as the ratio between the speech energy and the background noise energy such that the latter is calculated in frames where speech is present. The transients are normalized to have the same maximal amplitude as speech. This is a common setup rather than defining a signal to transient ratio due to the short duration of the transients [16].

The simulated signals are sampled at 16 kHz and are processed in consecutive time frames of length $T = 32$ ms (512 samples) with 50% overlap. The speech utterances used in the experiments are taken from TIMIT database [49]. The training set is composed of 20 speech utterances, and the test set, is composed of different 30 speech utterances. Each utterance is approximately of 9 s long and following the experimental setup in [16] is composed of three parts. The first part contains speech and background noise (without transients), the second part contains background noise and transients (without speech) and the third part contains the all three signals- speech, background noise and transients.

For the implementation of the proposed algorithm, we use the scattering transform library available in [50]. We exploit the Morlet wavelet similarly to [1] and set the quality factor to a small value $Q = 1$ for both the first and the second orders of the transform. Note that this choice of the quality factor Q provide filters with a small time support and they better characterize transients which are assumed to be of a short length. In addition, we use filters with a central frequency $\lambda > 2\pi/T$ such that the filter bank $\{\psi_\lambda\}_\lambda$ adequately covers the frequency axis. The number of the coefficients of a single frame for this setting is 9 and 36 for the first and the second orders of the scattering transform, respectively. The normalization parameter in (4.7) is set to $\epsilon = 3$, as was proposed in [16]. For the voice activity measure, the hyperplane of the SVM is optimized using standard MATLAB software using a Gaussian kernel with a variance $\sigma^2 = 1$ and the soft margin parameter is set to 1. We remark that these parameters are set to the default values of the software, they may be further optimized using a validation set to improve the classification results, and their optimization is not in the scope of this study. In addition, we empirically set the smoothing parameter in (4.10) to $J = 2$, which induce a lag of 32 ms.

Both for training the algorithm and for evaluating its performance on the test set, a ground truth is set according to the clean speech signal. A frame is considered as a speech frame if the energy of the clean signal in the frame is above a certain threshold $\tilde{\alpha}$. Namely, the speech indicator defined in (4.2) is given by:

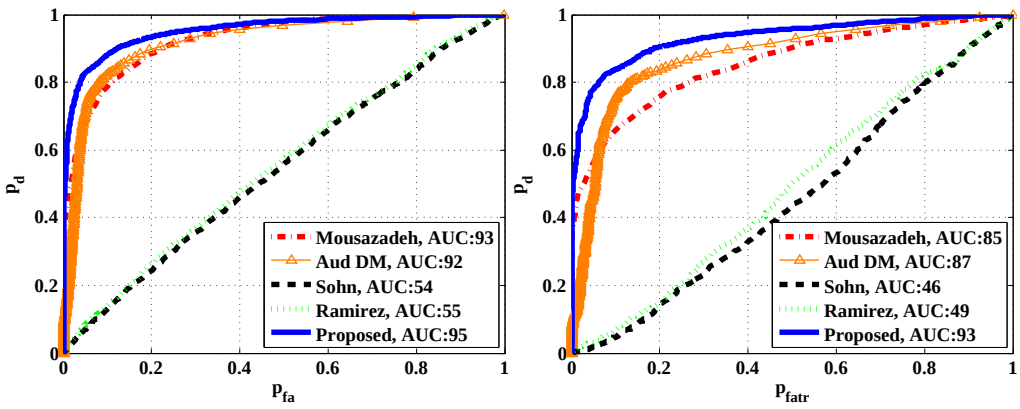
$$\mathbb{1}_{\mathbf{s}}(t) = \left\{ \begin{array}{ccc} 1 & ; & ||x_t||^2 > \tilde{\alpha} \\ 0 & ; & \text{otherwise} \end{array} \right\} \tag{4.12}$$

where x_t is the clean speech signal in frame t. The threshold $\tilde{\alpha}$ is set as the maximal threshold such that thresholding the speech signal has negligible auditory effect [16]. This setting of the ground truth is adequate for applications, e.g., speech recognition, where the detection of voice have a fine resolution. In this context we note that the algorithm presented in Chapter 3 is designed to operate in applications where voice is detected with lower resolution, e.g. for dominant speaker identification [3], where frequent switching between speech and non-speech hypotheses should be avoided. Therefore, it is adapted for the current setting of the ground truth by averaging the activity measures P^S in (3.21) and P^{US} in 3.22 over shorter time periods, i.e, by setting $L^S = L^{US} = 2$ in (3.21) and 3.22.

The performance of the algorithms is evaluated in the form of ROC curves, i.e., plots of probability of detection versus the probability of false alarms. The ROC curves are generated by sweeping the threshold over all possible values of the voice activity measure P_t in (4.10). We use two types of probabilities of false alarm as in [16]. The first is denoted by $\mathrm{P_{fa}}$ and is defined as the probability that a non-speech frame (which may contain a transient or may not) is detected as a speech frame. The second is denoted by $\mathrm{P_{fatr}}$ and is defined as the probability that a non-speech frame that contains a transient is wrongly detected as a speech frame. Namely, $\mathrm{P_{fa}}$ allows for evaluating the general performance of the algorithms, while $\mathrm{P_{fatr}}$ provides an insight on the performance of the algorithms in frames where transients are present. Note that the ground truth for the transients which is used for the evaluation of $\mathrm{P_{fatr}}$ is set in a similar way to the speech presence ground truth in (4.12). In addition, we evaluate the performance of the algorithm in terms of the Area Under the Curve (AUC) score, which is a scalar measure given by integrating the probability of detection over all values of false alarms. The AUC score of each method is given in percents in the legend box of each plot, and the higher the AUC the better the performance of the algorithm.

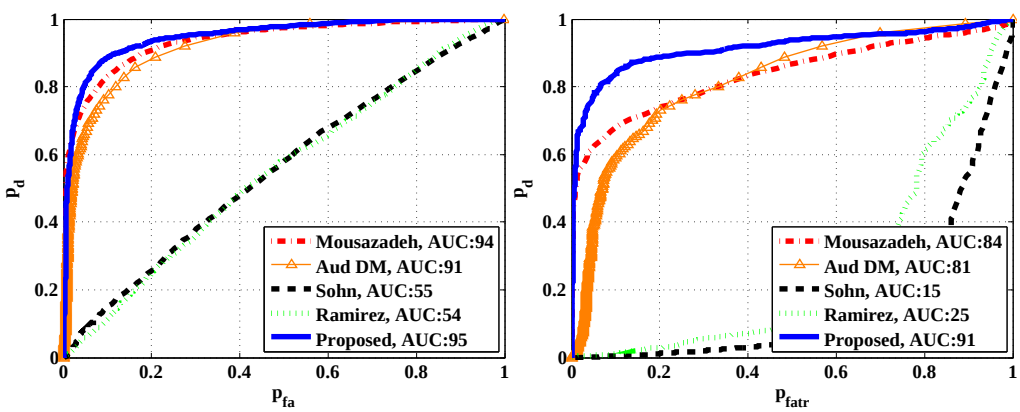
The experimental results are presented in Figures 4.1 to 4.3. It can be seen that both the proposed method, the audio version of the algorithm presented in Chapter 3, and the method presented in [16], which are specifically designed to perform in a highly non-stationary acoustic environment, significantly outperform the methods presented in [5] and [6]. Although, the audio version of the VAD presented in Chapter 3 is designed to

operate in applications where frequent swithcing between speech and non-speech hypotheses should be avoided, it still performs well also in voice activity detection in fine time scales, as in the experiments in this chapter. The proposed method, which is based on the scattering transform, provides improved results comparing to the competing methods, and in particular provides higher classification results in the presence of transients as demonstrated by the plots with the second type of false alarm P_{fatr}.



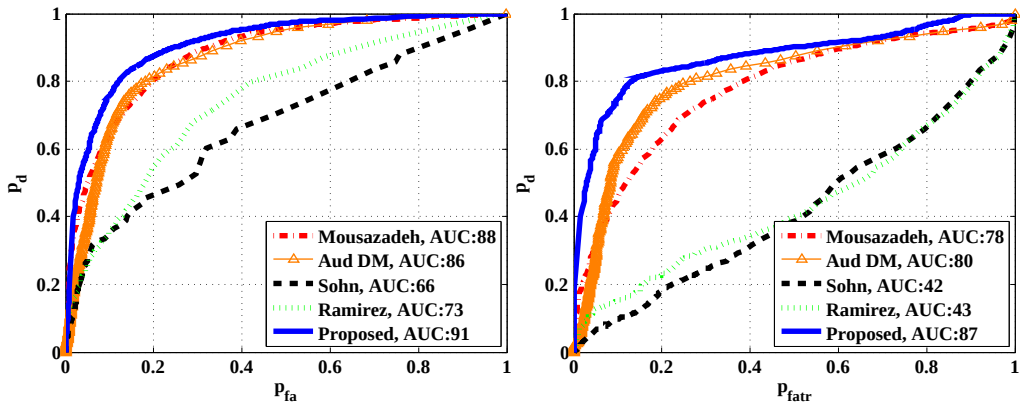
(a) (b)

Figure 4.1: (a) Probability of detection vs probability of false alarm (P_{fa}), and (b) Probability of detection vs probability of false alarm in transient frames (P_{fatr}). Test for a Gaussian noise with 0 dB SNR and keyboard typing transients.



(a) (b)

Figure 4.2: (a) Probability of detection vs probability of false alarm (P_{fa}), and (b) Probability of detection vs probability of false alarm in transient frames (P_{fatr}). Test for a colored Gaussian noise with 0 dB SNR and scissors transients.



(a) (b)

Figure 4.3: (a) Probability of detection vs probability of false alarm (P_{fa}), and (b) Probability of detection vs probability of false alarm in transient frames (P_{fatr}). Test for a babble noise with 0 dB SNR and door knocks transients.

4.5 Conclusions

We have presented a supervised learning algorithm for voice activity detection. The algorithm incorporates features extraction procedure, where the features are based on the scattering transform, and allow for a good separation between speech frames and non-speech frames which contain transients. In addition, the features incorporate noise estimation procedure and low weights are assigned to non-speech frames which contain background noise, separating them from the speech frames. The features are used to define a continuous measure for voice activity based on the SVM classifier. The proposed algorithm outperforms state-of-the-art VADs and in particular provides enhanced voice activity detection in presence of transients.

Chapter 5

Conclusion

5.1 Summary

In this research we have addressed the problem of voice activity detection in the presence of highly non-stationary noise and transients, and have presented two supervised learning algorithms for voice activity detection. These algorithms rely on different types of representation of the speech signal, i.e., diffusion maps and the scattering transform, which are designed to separate speech and non speech frames. The first algorithm exploits the video signal which is completely invariant to the acoustic environment along with the audio signal. First, the signals represented in a features domain which is based on MFCCs and incorporate noise estimation for the audio signal, and on motion vectors for the video signals. Then, diffusion maps is applied in the features domain of each modality and provides a parametrization of the data, i.e., a new representation of the signals which incorporates the advantages of the features and the diffusion mapping. The new representation is compact, is calculated with a low computational cost using the extension procedure, robust to noise, and separates speech and non-speech frames. In addition, this representation is unitless, and as a result, allows for merging data captured by different types of sensors. Using the new representation, we defined a measure of voice activity which incorporates a supervised learning procedure, i.e., GMM, and an unsupervised procedure, where large diffusion distance between consecutive frames indicate on voice activity. The voice activity measures of each modality are finally merged into a bimodal measure, and the speech indicator is estimated by comparing the bimodal measure to a

threshold. The experimental results demonstrate improved performance of single modal versions of the proposed VAD compared to state-of-the-art detectors. In addition, the proposed audio-visual VAD outperforms each of the single modal versions demonstrating the advantages of bimodal processing.

The second algorithm relies only on the audio signal and is based on the representation of the speech signal with the scattering transform. The scattering transform is constructed through a set of convolutions with filter banks and modulus operators. In this research we exploited the coefficients of the first two orders of the scattering transform as feature vectors which separates well speech and transient frames. These features are weighted with a scalar which incorporates noise estimation and receives low values in frames where only background noise is present. As a result, the proposed representation allows for a good separation between speech frames and both noise and transient frames. The features are used to define a new measure of voice activity which is based on the distance of a tested frame to a hyperplane, where the latter is optimized, similarly to the SVM procedure, to separate speech and non-speech clusters using labeled training data. Finally, the speech indicator is estimated by comparing the measure to a threshold. The proposed algorithm significantly outperforms state-of-the-art VADs which are based on statistical model. In addition, improved performance are demonstrated compared to the recently presented VAD in [16], in particular in the classification of frames which contain transients.

5.2 Future Research

The work presented in this thesis opens a number of research directions that can be further investigated:

1. In the present study, equal weights are assigned in the merging scheme of the two modalities. In future research, adaptive merging schemes may be developed, which incorporate estimates of the quality of the signals. For example, the weighting parameter may be adapted over time according to an estimate of the SNR of the audio signal, such that low values are assigned to the parameter for low values of the SNR, and the classification is mainly based on the video signal. Another example is using only the audio signal for the classification when the speaker moves his head

out of the frame, making the video signal irrelevant. Eye tracking algorithm may be applied to detect such scenarios.

2. To further improve the separation between speech and transients, an estimate of the transients may be incorporated in the representation of the audio signal. For example, an estimate of the spectral variance of the transients (e.g., as proposed in [15]) may be used along with the estimate of the spectral variance of the noise for the calculation of the scalar which weights the features of the audio signal in the both proposed algorithms. In this case, low weights would be assigned to the features both in the presence of background noise and transients, which may improve the separation between speech and transient frames in the features domain.

3. We have presented two types of representations of speech signals in highly non-stationary environments. The first representation is based on diffusion maps, the second is based on the scattering transform and both incorporate noise estimation. These representations of the speech signal may be further used for other applications rather than voice activity detection, e.g., speech recognition and dominant speaker identification in noisy environments [3].

4. In the present study we found that the scattering transform may be successfully used for the representation of speech signals in a highly non-stationary environment, and it allows for a good separation between speech and transients. However, the use of high orders of the scattering transform significantly increase the dimensionality of the data. Therefore, applying dimensionality reduction methods, such as diffusion maps, to the scattering features, may provide a representation which is both compact and benefit from the advantages of the scattering transform.

5. Diffusion maps are applied in this study, similarly and separately to the features of the audio and the video signals. Therefore, diffusion maps separately reveal the underlying structure of each modality. One research direction is to develop data driven tools which are based on similar concepts to diffusion maps but reveal only the mutual underlying structure of signals which are captured by different types of sensors. For example in audio-visual speech signals, these tools would reveal the

structure of speech and not the transient since the latter is captured only by the microphone but is not captured by the camera.

Bibliography

[1] J. Anden and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, Aug 2014.

[2] 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.

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

[4] A. Benyassine, E. Shlomot, H. Su, D. Massaloux, C. Lamblin, and J. Petit, "Itu-t recommendation g. 729 annex b: a silence compression scheme for use with g. 729 optimized for v. 70 digital simultaneous voice and data applications," *IEEE Communications Magazine*, vol. 35, no. 9, pp. 64–73, 1997.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

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

[13] ——, "Spectral enhancement by tracking speech presence probability in subbands," in *Proc. International Workshop on Hands-Free Speech Communication*, 2001.

[14] 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.

[15] A. Hirszhorn, D. Dov, R. Talmon, and I. Cohen, "Transient interference suppression in speech signals based on the OM-LSA algorithm," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

[16] 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.

[17] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[18] 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.

[19] 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.

[20] R. Talmon, I. Cohen, and S. Gannot, "Single-channel transient interference suppression with diffusion maps," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 132–144, 2013.

[21] 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.

[22] B. Rivet, L. Girin, and C. Jutten, "Visual voice activity detection as a help for speech source separation from convolutive mixtures," *Speech Communication*, vol. 49, no. 7, pp. 667–677, 2007.

[23] 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.*

[24] D. Scott, C. Jung, J. Bins, A. Said, and A. Kalker, "Video based VAD using adaptive color information," in *Proc. 11th IEEE International Symposium on Multimedia (ISM)*, 2009, pp. 80–87.

[25] C. Lopes, A. Goncalves, J. Scharcanski, and C. Jung, "Color-based lips extraction applied to voice activity detection," in *Proc. 8th IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 1057–1060.

[26] 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.

[27] 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.

[28] 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.

[29] V. Minotto, C. Lopes, J. Scharcanski, C. Jung, and B. Lee, "Audiovisual 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.

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

[31] 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.

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

[33] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *Journal of Shanghai University (English Edition)*, vol. 8, no. 4, pp. 406–424, 2004.

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

[35] 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.

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

[37] 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.

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

[39] C. Baugé, M. Lagrange, J. Andén, and S. Mallat, "Representing environmental sounds using the separable scattering transform," in *Proc. 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8667–8671.

[40] 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.

[41] 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.

[42] 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.

[43] J. Bilmes *et al.*, "A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," *International Computer Science Institute*, vol. 4, no. 510, p. 126, 1998.

[44] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[45] P. S. Aleksic and A. Katsaggelos, "Audio-visual biometrics," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 2025–2044, 2006.

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

[47] 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.

[48] S. Mallat, "Group invariant scattering," *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, 2012.

[49] 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.

[50] *[Online]. Available: http://www.cmap.polytechnique.fr/scattering.*

עיבוד אותות דיבור משולב שמע- וידאו
באמצעות מפות דיפוזיה והתמרת פיזור

דוד דב

עיבוד אותות דיבור משולב שמע- וידאו
באמצעות מפות דיפוזיה והתמרת פיזור

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת תואר מגיסטר למדעים

בהנדסת חשמל

דוד דב

הוגש לסנט הטכניון – מכון טכנולוגי לישראל

כסליו תשע"ה חיפה נובמבר 2014

תודות

המחקר נעשה בהנחיית פרופ' ישראל כהן מהפקולטה להנדסת חשמל בטכניון.

אני רוצה להביע את הערכתי לפרופ' ישראל כהן על ההנחיה, ההדרכה והתמיכה במהלך המחקר.

תודה לפרופ' רונן טלמון על שיתוף הפעולה והעזרה הגדולה במהלך המחקר.

תודה רבה למשפחתי ולחברי, שתמכו בי לאורך הדרך ובפרט התנדבו לעזור לי בהקלטות בהן עשינו שימוש במהלך המחקר.

לבסוף, תודה גדולה לאשתי חן ולבני עידן על האהבה והתמיכה האינסופיים.

מחקר זה נתמך על ידי הקרן הלאומית למדע (מענק מס' 1130/11)

אני מודה לטכניון על התמיכה הכספית הנדיבה בהשתלמותי.

תקציר

גילוי דיבור הוא רכיב יסודי ביישומים רבים בתחום עיבוד הדיבור כגון זיהוי דיבור, זיהוי דוברים, קידוד דיבור, הורדת רעשים וזיהוי דובר דומיננטי. שיטות מוקדמות לגילוי דיבור מבוססות על תכונות פשוטות כגון האנרגיה של האות וקצב חציית האפס. למרות ששיטות אלו פועלות בצורה טובה עבור אותות נקיים, הביצועים שלהם מתדרדרים בנוכחות רעשי רקע מפני שלדוגמא, האנרגיה של האות הנקלט עלולה לגדול עקב רעש הרקע ולהצביע באופן שגוי על נוכחות דיבור. על מנת להתגבר על בעיה זו, הוצגו בשנים האחרונות מספר גלאי דיבור שמניחים מודלים סטטיסטיים על האות והרעש, ומבוססים על מבחן הסבירות המרבית. מבין המודלים הסטטיסטיים ניתן למצוא את המודל הגאוסי, המודל הלפלסי ומודל גאמה המוכלל. מבחן הסבירות המרבית, בשיטות אלו, כולל שיערוך של השונות הספקטרלית של הרעש, ושיטות נפוצות לשערוך זה מניחות, בדרך כלל, כי הרעש משתנה לאט ביחס לאות. הנחה זו אינה תקפה לרעשים שאינם סטציונריים ולטרנזיאנטים שהם הפרעות קצרות זמן כגון הקשות מקלדת ודפיקות בדלת, והביצועים של שיטות אלו מתדרדרים באופן משמעותי.

באחרונה, הוצג אלגוריתם המבוסס על למידה מפוקחת לגילוי דיבור בנוכחות רעשים לא סטציונריים. האלגוריתם כולל חישוב וקטורי-תכונות שמבוססים על מקדמי סקלת מל (Mel-Frequency Cepstral Coefficients) וכוללים שערוך של הרעש. וקטורי התכונה משמשים לביצוע אשכול (clustering) ספקטרלי לצורך גילוי דיבור. למרות שדווח על ביצועים טובים של האלגוריתם ביחס לאלגוריתמים עדכניים, הביצועים שלו עדיין מוגבלים בנוכחות של טרנזיאנטים, מפני שדיבור וטרנזיאנטים מופרדים רק באופן חלקי במרחב וקטורי התכונה. לכן, גילוי דיבור בנוכחות רעש לא סטציונרי וטרנזיאנטים נשאר בעיה פתוחה.

אחד הפתרונות לבעיה הוא לשלב מידע חזותי, כלומר צילום וידאו של הדובר, שהוא בלתי תלוי בסביבה האקוסטית והוא זמין במגוון רחב של מוצרים, כגון חדרי ועידת וידאו, שירותים חינמיים של שיחות וידאו על גבי רשת האינטרנט ובאמצעות מכשירים סלולריים חכמים. לכן, שימוש במידע חזותי יכול להיות מועיל ופרקטי לצורך גילוי דיבור בסביבות אקוסטיות מאתגרות, ואכן, מספר אלגוריתמים לגילוי דיבור שמנצלים את אות הווידאו הוצגו בספרות בשנים האחרונות. שיטות אלו בדרך כלל מבוססות על צילום וידאו של אזור הפה של הדובר ועקיבה אחרי הצורה של הפה והדינמיקה שלו. למרות שבשיטות אלו משתמשים בוקטורי תכונה שונים, ההנחה שבדרך כלל חוזרת על עצמה היא שפה פתוח

מזוהה עם דיבור. מכיוון שייתכנו מסגרות זמן בהן הפה סגור במהלך הדיבור, מניחים בנוסף כי הדיבור מאופיין בתנועה גדולה של השפתיים. בעוד שלגלאי דיבור המבוססים על אות הווידאו קיימים יתרונות בסביבות אקוסטיות רועשות, הם אינם יכולים להתחרות בגלאי דיבור המבוססים על אות השמע בסביבות נקיות או כאלו בהן רמת הרעש הסביבתי היא נמוכה. זאת מפני שאפילו מאפייני דיבור פשוטים, כגון האנרגיה של האות, מספקים ביצועים כמעט מושלמים בסביבה אקוסטית שקטה. לכן, השילוב של אות הווידאו יחד עם אות השמע יכול להיות מועיל לצורך גילוי דיבור. למרות זאת, רק אלגוריתמים בודדים שמשלבים את אותות השמע והווידאו לצורך גילוי דיבור קיימים בספרות כיום.

בעבודת מחקר זו אנחנו מציגים שני אלגוריתמים לגילוי דיבור המבוססים על למידה מפוקחת, ואנחנו מניחים כי קיים סט אימון מסומן הכולל מקטעי דיבור בסביבה של רעשי רקע וטרנזיאנטים. באלגוריתם הראשון, אנחנו מניחים כי אות הדיבור נקלט על ידי שני חיישנים שונים, מיקרופון ומצלמה, כאשר צילום הווידאו כולל את אזור הפה של הדובר. האלגוריתם מבוסס על ההנחה שלדיבור יש מבנה גאומטרי יסודי הנובע מהפיסיקה של היווצרותו ואינו תלוי בחיישנים שבהם הוא נקלט. על מנת לחשוף את המבנה היסודי, אנחנו מפעילים מיפוי דיפוזיה, שהוא שיטה גאומטרית מבוססת מידע, בצורה זהה ונפרדת לכל אחד מהאותות. מיפוי הדיפוזיה מופעל במישור של וקטורי תכונה המתוכננים להפריד בין מסגרות זמן שכוללות דיבור לכאלו שלא. וקטורי התכונה של אות השמע מבוססים על מקדמי סקלת מל וכוללים משקולות שמבוססות על שערוך רעש רקע, כך שבמסגרות זמן בהן קיים רק רעש רקע יתקבלו משקולות נמוכות. וקטורי התכונה של אות הווידאו מבוססים על וקטורי תנועה ומאפשרים לאפיין את הצורה של הפה והדינמיקה שלו. לצורך המימוש של מיפוי הדיפוזיה, נדרש סט של מסגרות זמן, ומכיוון שהאלגוריתם המוצע נדרש להיות מתאים לעיבוד האות בזמן אמת, בפועל, המימוש מתבצע על סט האימון. בהינתן מסגרת זמן בסט המבחן, חישוב מיפוי הדיפוזיה שלה מתבצע על ידי אינטרפולציה ביחס למסגרות זמן דומות בסט האימון. תהליך האינטרפולציה מאפשר חישוב מדויק ויעיל חישובית של מיפוי הדיפוזיה והופך את האלגוריתם מתאים לעיבוד האות בזמן אמת. הפעלת מיפוי הדיפוזיה במישור וקטורי התנועה מספק ייצוג מממד נמוך של האותות. אנחנו משתמשים בייצוג החדש כדי להגדיר מדד לגילוי דיבור הכולל פרוצדורת למידה מפוקחת המבוססות על מודל תמהיל גאוסי (Gaussian Mixture Model), וכן פרוצדורת למידה שאינה מפוקחת שמניחה קצבי שינוי שונים במהלך מקטעי זמן שכוללים דיבור וכאלו שלא. מדדי הדיבור שמבוססים על אותות השמע והווידאו מאוחדים בסופו של דבר למדד יחיד שמשמש לגילוי דיבור. תוצאות הניסויים מדגימים את הביצועים המשופרים של האלגוריתם המוצע ביחס לאלגוריתמים עדכניים. בנוסף, אנחנו מדגימים את היתרון שבשילוב בין אותות השמע והווידאו ביחס לגרסאות של האלגוריתם המוצע המבוססות על מודליות (modality) יחידה לצורך גילוי דיבור.

באלגוריתם השני אנחנו מתמקדים על אות השמע והוא מבוסס על התמרת פיזור. התמרת הפיזור מתקבלת על ידי אוסף של קונבולוציות של האות המורעש עם בנק של פילטרים

ואופרטורי ערך מוחלט, והיא מספקת ייצוג של האות בצורה של רשת עמוקה. ייצוג האות על ידי הסדר הראשון של התמרת הפיזור דומה לייצוג האות על ידי מקדמי סקלת מל. אולם, שני ייצוגים אלו סובלים מאיבוד מידע על פני סקלות זמן קצרות הנובע מהחלקה של האות. לדוגמא, ייצוג של טרנזיאנטים שהם קצרים בזמן, יכול להיות דומה לייצוג של דיבור עקב ההחלקה ולהוביל לגילויי דיבור שגויים. ניתן לראות בהתמרת פיזור כהרחבה של מקדמי סקלת מל כך שסדרים גבוהים של ההתמרה משחזרים את המידע שאבד, מה שהופך את התמרת פיזור למתאימה לייצוג של אותות דיבור בסביבה של טרנזיאנטים שהם קצרים בזמן. האלגוריתם המוצע משתמש בוקטורי תכונה המבוססים על התמרת פיזור ומאפשרים הפרדה בין דיבור וטרנזיאנטים, ובנוסף הוא כולל פרוצדורת שערוך רעש המאפשרת הפרדה בין דיבור לרעש הרקע. על ידי שימוש בוקטורי תכונה אלו, אנחנו מציעים מדד חדש לגילוי דיבור המבוסס על מסווג מכונת וקטורי תמיכה (Support Vector Machine). בניגוד למסווג הרגיל שמאפשר סיווג בינארי על סמך מיקום של וקטור תכונה נבחן ביחס למישור שמפריד בין שני אשכולות, המדד המוצע הוא רציף והוא מבוסס על ההנחה שככל שווקטור תכונה נבחן רחוק יותר מהמישור המפריד כך הודאות לסיווג נכון שלו גדולה יותר. גילוי דיבור מתבצע על ידי השוואת המדד הרציף לסף כך שאם המדד מעל סף מסוים אז קיים דיבור. היתרון בשימוש במדד רציף הוא שניתן לשלוט ברגישות הגילוי על ידי בחירת סף מתאים ליישום נתון. תוצאות הניסויים מדגימים כי האלגוריתם המוצע מספק ביצועים עדיפים על אלגוריתמים עדכניים ובפרט מאפשר גילוי דיבור מדויק במסגרות זמן שכוללות טרנזיאנטים.