Design and Analysis of a Constant Beamwidth Beamformer

Oren Rosen
Design and Analysis of a Constant Beamwidth Beamformer

Research Thesis

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

Oren Rosen

Submitted to the Senate of the Technion—Israel Institute of Technology
5776 Haifa February 2016
Acknowledgement

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 my advisor, Professor Israel Cohen for his guidance and support throughout all research stages.

I would like to thank Professor David Malah for his substantial aid in the writing of this thesis.

Finally, I would like to thank my father Avi and wife Natalie for their support and patience during all stages of my research.

This work is dedicated to the memory of my late mother, Ilana.
Contents

1 Introduction .................................................. 9
  1.1 Motivation and Overview .................................. 9
  1.2 Thesis Structure ......................................... 13

2 Related Work and Theoretical Background .................. 15
  2.1 Propagating Wave Fields ................................ 16
  2.2 Microphone Arrays ..................................... 18
  2.3 Time-Space Signals .................................... 19
  2.4 Microphone Array Performance Measures ................. 21
  2.5 Finite Impulse Response Filter Based Filtering ........ 22
  2.6 Constant Beamwidth Beamformers ........................ 24
    2.6.1 Multi Beamforming with FIR Filtering Approach .. 25
    2.6.2 Least Squares Approach ............................ 27
  2.7 Voice Activity Detection ................................ 31
  2.8 Transient Noises ..................................... 33
  2.9 Spectral Clustering ................................... 35
  2.10 Mel Frequency Cepstrum Coefficients ................. 36

3 Speech Enhancement Using Beamformers .................... 37
  3.1 Background and Problem Formulation .................... 37
    3.1.1 Microphone Array Configuration .................... 38
    3.1.2 Temporal Filtering by an FIR Filter Array ........ 40
  3.2 Proposed FIR Design Algorithm ........................ 43
List of Figures

1.1  A speech enhancement system integrating between the two proposed algorithms in this thesis ................................................................. 12

2.1  A three dimensional coordinate system with Cartesian and spherical coordinates. ................................................................. 16

2.2  A uniform linear array of size $M$ and spacing $d$. ......................................................... 19

2.3  A simple beamformer design. .......................................................... 20

2.4  FIR beamformer architecture, where D denotes a delay element. ................................. 23

3.1  A uniform linear array of size $M$: (a) $M$ even and (b) $M$ odd, with spacing $d$. The red dotted line denotes the symmetry point of the array located at $x = \left(\frac{M-1}{2}\right) d$. ................................................................. 39

3.2  The desired frequency responses of the FIR filters for the lower half of the frequency range, indexed by microphone index $m = 0, \ldots, \frac{M-1}{2}$, $M$ odd. Due to the microphone array symmetry, the magnitude response specification for the filters connected to the second half of the array, i.e., for $m = \frac{M+1}{2}, \ldots, M-1$, is a mirror image of first half of the array. Each dot in the specified response is the value of the filter magnitude at a corresponding frequency bin. Since $M$ is odd and due to symmetry, the two bottom rows of Fig. 3.2 represent all-pass FIR filters since $M_{\text{min}} = 3$. This means that three microphones are effective at all frequencies. For $M$ even, only two microphones would be effective at all frequencies since $M_{\text{min}} = 2$ and therefore only the FIR filter connected to the microphone indexed by $m = \frac{M-1}{2}$, i.e., the last row, would be an all-pass filter. ......................................................... 47
3.3 Array beampattern of (a) a conventional array without FIR filters. (b) an FIR array with magnitude response $W_L$, as given in (3.18). And (c) the modified design presented in the next section. The frequency range of $[0, \frac{1}{2T}] = [0,8]$ kHz is represented by the Y axis and is quantized into $N = 32$ frequency bins. The DOA range of $\theta = [-90,90]$° is represented by the X axis. For display purposes, the frequency response at a higher resolution is computed from the FIR filter coefficients. The desired beampattern beamwidth is $\theta_{BW} = 30°$ and the minimal frequency is $f_0 = 3412$ Hz.

3.4 Modified FIR magnitude responses for $M$ odd, where $W_{k,m}$ denotes the FIR filter magnitude response at the frequency bin indexed by $k$ and the microphone indexed by $m$.

3.5 Proposed FIR-based beamformer structure.

3.6 Simulation results for ULA frequency response of (a) the proposed algorithm. (b) Parra’s algorithm. (c) Tourbabin’s algorithm. (d) Doblinger’s algorithm.

3.7 Beampattern beamwidth error of the proposed algorithm (blue square). Parra’s algorithm (red circle). Tourbabin’s algorithm (green solid). Doblinger’s algorithm (black triangle). conventional ULA (cyan star).

3.8 Simulation results for ULA beampatterns at different frequencies of (a) the proposed algorithm. (b) Parra’s algorithm. (c) Tourbabin’s algorithm. (d) Doblinger’s algorithm. The side-lobe levels near the main-lobe are marked by the red dashed lines and the side-lobe levels at the edges, due to aliasing, are marked by the black dash-dot lines.


3.10 Beamwidth error between the microphone gain mismatched array and the original array of the proposed algorithm (blue square), Parra’s algorithm (red circle), Tourbabin’s algorithm (green solid) and Doblinger’s algorithm (black triangle).
3.11 Beamwidth error between the microphone spacing mismatched array and original array of the proposed algorithm (blue square), Parra’s algorithm (red circle), Tourbabin’s algorithm (green solid) and Doblinger’s algorithm (black triangle). .................................................. 64

4.1 A block scheme of the proposed (a) training, and (b) testing stages. ........ 71

4.2 Probability of detection ($P_d$) versus probability of false alarm ($P_{fa}$), for various noise environments. (a) Babble noise with SNR of 5dB, and transient noise of door knocks. (b) White noise with SNR of 5dB, and transient noise of door knocks. (c) White noise with SNR of 10dB, and transient noise of typing. (d) Babble noise with SNR of 5dB, and transient noise of typing. ............... 74

4.3 Clustering results of training stage for babble noise with SNR of 5dB, and transient noise of typing of (a) the proposed algorithm, and (b) the algorithm proposed in [44]. ................................................. 75
List of Tables

3.1 Proposed FIR based constant beamwidth beamformer algorithm . . . . . 56
3.2 Side-lobe attenuation of the various methods before and after the micro-
phone array gain perturbation. The percent of degradation for every
method is given in brackets. . . . . . . . . . . . . . . . . . . . . . . . . . . . 63
3.3 Side-lobe attenuation of the various methods before and after the microphone array spacings perturbation. The percent of degradation for every
method is given in brackets. . . . . . . . . . . . . . . . . . . . . . . . . . . . 65
4.1 Proposed Voice Activity Detection Algorithm Based on Spectral Clustering
Method. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 76
List of Publications


Abstract

Beamforming algorithms are widely used in many real world applications such as radar, sonars, medical diagnosis, teleconferencing and seismic sensing. They are utilized to solve several problems in the areas of signal processing and communications. To name a few, for example, they can provide significant spatial filtering by suppressing both interference signals and noise signals while perfectly recovering the desired signals. Beamformers can be also used for direction of arrival and localization applications, dereverberation, and more.

In the literature, several approaches were proposed for the design of a constant beamwidth beamformer. Some of these approaches suffer from performance degradation in some aspects like poor sidelobe attenuation, reduced interference signal attenuation and sensitivity to model mismatch. Other approaches suffer from high computational complexity which make them infeasible in some applications involving large arrays or strict hardware limitations. Although some design methods provide fairly robust beamformer performance, the challenge of designing a constant beamwidth beamformer with noise robustness and reduced sensitivity to parameters mismatch is still a challenging task.

In this thesis, we introduce a new approach for designing a constant beamwidth beamformer. The proposed approach utilizes custom-tailored finite impulse response filters for each microphone channel, manipulating the beampattern beamwidth. The manipulated beampattern has approximately a constant beamwidth over a wide frequency band. By exploiting the physical microphone array configuration and attributes, we shape accordingly the frequency response of the finite impulse response filters and control the beamformer beamwidth. The proposed approach demonstrates low computational complexity as well as improved array response results in various scenarios, compared to other methods in the field. This approach could be used for beam steering in teleconferencing or speaker track-
Beamformer steering is directing the beamformer’s main beam to a desired direction by adding appropriate delays or phase shifts. This ability is required especially in applications where not only the location of the source is needed but also the ability to know when the source was active, as in teleconferencing and automatic video tracking.

As a part of this thesis, we also develop a voice activity detection algorithm based on spectral clustering and diffusion kernels. The proposed method is based on obtaining a low dimensional representation from feature vectors, which are viewed as a cloud of points. Then, we incorporate this representation in a Gaussian Mixture Model in order to build a statistical model for labeling data as speech or non-speech. Simulation results demonstrate improved performance of the proposed algorithm compared to a recent VAD algorithm in presence of environmental noise. This approach could be useful for teleconferencing and speaker tracking devices as well to further improve the speaker localization. For example, a teleconferencing device may achieve improved energy efficiency by only processing sequences where speech is present.
Notation

Notations for Chapter 3

c velocity of sound propagation in air

d inter-microphone spacing

\( D(\omega, \theta) \) beampattern

\( f(t,x) \) set of signals sampled by the microphone

\( f_0 \) minimal frequency to achieve a desired beamwidth

\( f_h \) higher boundary of a selected frequency range

\( f_l \) lower boundary of a selected frequency range

\( f_{\text{max}} \) maximal frequency to achieve a desired beamwidth

\( f_p \) minimal frequency to achieve a desired beamwidth for \( M_p \) effective microphones

\( f_\omega(t,x) \) propagating plane wave

\( k \) bin index of the DFT

\( L \) number of frequency bins in the lower band

\( m \) microphone index

\( M \) number of microphones in an array

\( M_{\text{min}} \) minimal number of participating microphones in an array

\( M_p \) number of effective participating microphones in an array

\( N \) number of FIR filter coefficients

\( P(\omega, \theta) \) beamformer response

\( T_s \) delay between adjacent filter elements

\( v_s(\omega, \theta) \) stacked array manifold vector

\( v(\kappa) \) array manifold vector

\( w \) complex weighting vector

\( W \) filter array matrix

\( W_H \) magnitude response of \( W \) of the filters in the upper band

\( \tilde{w}_k \) samples of the normalization filter frequency response

\( W_L \) magnitude response of \( W \) of the filters in the lower band

\( w_m \) complex weighting of a microphone

\( w_{m,n}^* \) a complex coefficient of the FIR filter
\( W_m(\exp(j\omega T_s)) \)  DTFT of \( w_{m,n} \)

**\( w_s \)**  composite stacked weight vector

**\( x_m \)**  microphone positioning

**\( y(t) \)**  beamformer output

**\( \theta \)**  direction of arrival

**\( \theta_{BW} \)**  desired beamwidth

**\( \kappa \)**  wavenumber for plane waves in a locally homogeneous medium

**\( \lambda \)**  wavelength

**\( \tau_m \)**  propagation delay of an incoming signal

**\( \omega \)**  temporal frequency
Notations for Chapter 4

$C^\ell_t$ indicator matrix for a speech containing frame
$D$ diagonal matrix whose $i$-th diagonal element equals to $\sum_{j=1}^{N} W(i, j)$
$E$ expected value of a matrix
$f(\cdot; \mathcal{H}_0)$ probability density function corresponding to speech absence frames
$f(\cdot; \mathcal{H}_1)$ probability density function corresponding to speech presence frames
$\mathcal{H}_0$ speech absence hypotheses
$\mathcal{H}_1$ speech presence hypotheses
$i$ desired number of samples per frame
$j$ frame index
$k$ current iteration index
$K_m$ MFCC coefficients frequency bins
$K_s$ STFT coefficients frequency bins
$m$ number of MFCCs
$M$ size of moving window
$N$ number of signal frames
$S$ diagonal matrix made by $\sum \sum^T$ and $\sum^T \sum$
$t$ time frame
$T_h$ threshold of tradeoff between probability of detection and false alarm
$U_{\ell t}$ a matrix made of $D^{\ell-1/2} W^{\ell} D^{\ell-1/2}$
$\tilde{U}(t,:)$ new representation of the unlabeled data
$V$ orthogonal matrix, its columns are the eigenvectors of $\sum^T \sum$
$VA_{\ell t}$ decision rule for an unlabeled time frame
$w_M$ moving window of size $M$
$W_{\ell opt}^{\ell}$ similarity matrix of the $\ell$-th training sequence
$x_{sp}(n)$ speech signal
$x_{st}(n)$ additive interfering stationary noise signals
$x_{ts}(n)$ additive interfering transient noise signals
$X^{ji}_m$ size $m \times i$ matrix of MFCCs of the $j$-th frame
$y(n)$ VAD microphone input signal
\( \hat{y}_j(n - k) \)  generated sample cloud of the \( j \)-th frame (out of \( N \))

\( Y_m(t, k) \)  absolute value of the MFCC in a given time frame

\( Y_s(t, k) \)  absolute value of the STFT coefficients in a given time frame

\( \gamma_k(t) \)  a-posteriori SNR

\( \Gamma_{\text{train}}^t \)  likelihood ratio for a labeled frame \( t \)

\( \Delta \)  orthogonal matrix of size \( 3 \times N \)

\( \epsilon \)  kernel width obtained during the training stage

\( \theta \)  a vector of system parameters

\( \Lambda_t \)  arithmetic mean of the log-likelihood ratios for frame \( t \)

\( \lambda_s(t, k) \)  variance of speech signal in the \( k \)-th frequency bin of the \( t \)-th frame

\( \lambda_n(t, k) \)  variance of stationary noise in \( t \)-th time frame and \( k \)-th frequency bin

\( \Lambda^t \)  arithmetic mean of the log-likelihood ratio

\( \Sigma_i^t \)  covariance matrix of the \( i \)-th frame in the \( \ell \)-th sequence

\( \Sigma \)  covariance matrix for the \( j \)-th frame

\( \Upsilon \)  approximate orthonormal basis of \( D^{-1/2}WD^{-1/2} \)

\( \xi_k(t) \)  a-priori SNR
Abbreviations

DOA  Direction of arrival
DFT  Discrete Fourier transform
DTFT Discrete-time Fourier transform
FIR  Finite impulse response
GMM Gaussian mixture model
GSM  Global system for mobile communication
LCMV Linearly constrained minimum variance
LRT  Likelihood Ratio Test
LS   Least squares
MAE  Mean absolute error
MCRA Minima Controlled Recursive Averaging
MFCC Mel-frequency cepstrum coefficient
MMSE Minimum mean square error
MVDR Minimum variance distortion-less response
SNR  Signal-to-noise ratio
SV   Spatial variation
SVD  Singular vector decomposition
TASI Time assignment speech interpolation
ULA  Uniform linear array
VAD  Voice activity detection
Chapter 1

Introduction

1.1 Motivation and Overview

Speech enhancement is used in many applications in communication and signals processing such as mobile devices, hearing aids, speech recognition and conferencing equipment [1–3]. Speech enhancement refers to the process of quality improvement of a speech signal by using digital signal processing algorithms. The main goal of speech enhancement is the improvement in intelligibility, a measure of comprehensibility assessment of the signal to the human ears [4,5]. The main difficulties in implementing effective speech enhancement systems is the presence of additive acoustic noise, especially noticeable environmental noise. Traditional speech enhancement methods are based on simple features such as signal energy levels and zero-crossing rate [6–8]. These standards demonstrate fairly good performance and are therefore widely used in communication. Nevertheless, their performances degrade in presence of environmental noise, even for relatively high SNR values.

One of the most prominent methods introduced to overcome this problem is the presence of directional noise, i.e., noise originating from various ambient sources. Directional noise may cause speech signals not to be acquired and processed in a clean, desirable manner. Similarly to other environmental noises, directional noise degrades speech quality and intelligibility. Thus, there is a need to clean the microphone signal and spatially filter it before it is processed. One of the possible ways to obtain spatial filtering is the utilization of directional microphones, i.e., microphones that their mechanical and electrical charac-
teristics enable enhancing signals from specific directions while suppressing signals from other undesired directions, such as the cardioid and shut-gun microphones [9, 10]. Other traditional speech enhancement methods utilize a single, omnidirectional microphone. In the single-channel scenario, the microphone input is modeled as a superposition of the speech and noise signals. The main drawback of these methods is that they distort the desired speech signal. Thus, it has been proposed to utilize microphone arrays in order to tackle this problem [11, 12]. These methods are also referred to as beamforming, since the main-lobe (beam) is directed at the desired location. Basic beamforming techniques such as the delay and sum [13] utilize a single weight for the entire frequency spectrum. This makes them effective only for narrowband signals and ineffective when dealing with wideband signals that are composed of various frequency components. To overcome this problem, constant beamwidth beamformers have been introduced. Constant beamwidth beamformers are beamformers that have a constant main-lobe width over a range of frequencies in a wideband. A constant beamwidth is desirable in wideband applications such as speech processing with microphone arrays and multiple user conferencing. Traditional beamformer design approaches did not provide sufficient performance in terms of beamwidth constancy and other performance criteria. Therefore, numerous contemporary approaches to robust beamformer designs were proposed. The minimum variance distortion-less response (MVDR) [14,15], linearly constrained minimum variance (LCMV) [16] and least-squares [17,18] are several widely used beamforming design techniques developed for improving the traditional methods. The least squares (LS) approach is a popular design method [18]. It can provide a closed-form solution to the problem and is more computationally efficient compared with the convex optimization method. Nevertheless, most contemporary LS methods are utilized for the design of general broadband beamformers and not necessarily for designing constant beamwidth beamformers. The task of beamforming and constant beamwidth beamforming in particular remains unsolved as beamformer design is based on balancing trade-offs between various performance criteria and physical constraints. One of the popular beamformer design approaches is the FIR based beamformer [19–23]. FIR beamformers perform temporal filtering that results in a frequency dependent response. The filtering is performed in order to compensate for the phase difference of various frequency components. It has been shown [24,25] that
utilizing narrowband FIR filters in beamformer designs can produce robust beamformers.

Another available solution for the environmental noise problem is the Voice Activity Detection (VAD). Classical VAD algorithms presuppose statistical models of the speech and noise signals, and many are based on Likelihood Ratio Test (LRT). Two of the most widely used statistical models are the Gaussian and Laplacian models [26, 27], where the LRT fuses estimations of the spectral variance of the noise, together with common methods for its estimation. These methods usually assume that the statistics of noise is slowly varying with respect to speech [28,29]. This assumption is invalid for non-stationary noise and transient interferences, which are abrupt bursts of sound and vary faster than speech such as coughing, sneezing, door knocking and keyboard typing [30,31]. Another difficulty is the presence of directional interference. Directional interference might be originated in acoustic reverberation, resulting from multiple reflections of acoustic signals, and in several signal origins, generating multiple signals from various directions. Finally, speech is often modelled in order to separate it from unwanted interference. These models are usually based on assumption that are incorrect in many cases of environmental and directional interference causing deteriorated performance in these cases. In the method recently presented in [32], the arithmetic mean of the LRT is used to weight features base mainly on the Mel-Frequency Cepstrum Coefficients (MFCC). 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.

The solution to the problem of speech enhancement in presence of environmental noise could be integration of the two aforementioned methods. By building a system based on two stages, the first, spatial filtering via utilization of a microphone array. By utilizing a beamformer, we enhance the speech signal impinging the array from the desired direction while suppressing all other undesired sources originating in undesired directions. This achieves a cleaner signal, comparing to the single-microphone scenario [33]. The second stage would be a VAD algorithm, cascaded to the output of the beamformer. After the incoming signal is spatially filtered, we could utilize the VAD algorithm for suppression of environmental noises that were not filtered in the beamforming stage. In particular,
CHAPTER 1. INTRODUCTION

non-stationary and transient noise, since beamformers do not entirely cancel these noise signals. The output of this cascaded system would yield a system much more robust to environmental noise, comparing to the performance of every method separately. The block diagram of the cascaded speech enhancement system is shown in Fig. 1.1.

Figure 1.1: A speech enhancement system integrating between the two proposed algorithms in this thesis

Another possible application for the proposed methods would be in beamsteering. In acoustic environments, the source localization is crucial for many applications such as automatic beamformer steering for suppressing noise and reverberation. Beamformer steering is directing the beamformer’s main beam to a desired direction by adding appropriate delays or phase shifts. This ability is necessary especially in applications where not only the location of the source is needed but also the ability to know when the source was active, as in teleconferencing and automatic video tracking. In the future, these two approaches could be integrated to work as a system, enabling a steerable constant beamwidth beamformer for teleconferencing or speaker tracking devices. Such a system would be able to locate the main speaker and determine when they have spoken, suppressing environmental noises and unwanted interruptions as well as enhance the energy efficiency by only working while the main speaker is talking.

In this thesis we present two speech enhancement components, the first is a microphone array design method for implementing a constant beamwidth beamformer. Custom-
CHAPTER 1. INTRODUCTION

tailored finite impulse response (FIR) filters are utilized for each microphone channel, manipulating the beampatterns beamwidths. The manipulated beampatterns have a constant beamwidth over a range of frequencies in a wideband. Additionally, a post summing output normalization filter is used to ensure a frequency invariant gain of the beampattern. By exploiting the physical microphone array configuration and attributes, we shape accordingly the frequency response of the FIR filters and eventually control the beamformer beamwidth. The proposed approach demonstrates improved array response results in various scenarios, compared to available methods in the field, especially in terms of sensitivity to parameters mismatch, noise robustness and side-lobe attenuation.

The second component is a supervised learning algorithm for voice activity detection transient noise suppression comprised of training and testing stages. A sample cloud is produced for every signal frame by utilizing a moving window. MFCCs are then calculated for every sample in the cloud, producing an MFCC matrix and subsequently a covariance matrix for every frame. Utilizing the covariance matrix, we calculate a similarity matrix using spectral clustering and diffusion kernels methods. Using the similarity matrix, we cluster the data and transform it to a new space where each point is labeled as speech or non-speech. We then use a Gaussian Mixture Model (GMM) in order to build a statistical model for labeling data as speech or non-speech. Simulation results demonstrate its advantages compared to a recent VAD algorithm.

1.2 Thesis Structure

The rest of this thesis is organized as follows: In Chapter 2, we review the theoretical background and present methods in the fields of array processing and voice activity detection. We start with the relevant basic theory for microphone array design. Then, we present the theory of wideband microphone array design. Specifically, we review FIR filter based array design methods. The main concept of our array processing research, which is the constant beamwidth beamformer is presented, together with two design methods in the field. We then introduce the relevant background theory and basics of VAD algorithms. We elaborate on one of the major obstacles in VAD, transient noises. We then introduce the theory behind spectral clustering, a common method for data analysis which
is the method used for the proposed VAD algorithm. Finally, we elaborate on MFCCs, that are the representation of the sound spectrum in the proposed VAD method.

In Chapter 3, we introduce a new approach for obtaining a constant beamwidth beamformer. The proposed approach utilizes custom-tailored finite impulse response (FIR) filters for each microphone channel, manipulating the beampatterns beamwidths. The manipulated beampatterns have a constant beamwidth over a range of frequencies in a wideband. Additionally, a post summing output normalization filter is used to ensure a frequency invariant gain of the beampattern. By exploiting the physical microphone array configuration and attributes, we shape accordingly the frequency response of the FIR filters and eventually control the beamformer beamwidth. The proposed approach demonstrates improved array response results in various scenarios, compared to available methods in the field, especially in terms of sensitivity to parameters mismatch, noise robustness and side-lobe attenuation.

In Chapter 4, we introduce a voice activity detection (VAD) algorithm based on spectral clustering and diffusion kernels. The proposed algorithm is a supervised learning algorithm comprising of learning and testing stages: A sample cloud is produced for every signal frame by utilizing a moving window. Mel-frequency cepstrum coefficients (MFCCs) are then calculated for every sample in the cloud in order to produce an MFCC matrix and subsequently a covariance matrix for every frame. Utilizing the covariance matrix, we calculate a similarity matrix using spectral clustering and diffusion kernels methods. Using the similarity matrix, we cluster the data and transform it to a new space where each point is labeled as speech or non-speech. We then use a Gaussian Mixture Model (GMM) in order to build a statistical model for labeling data as speech or non-speech. Simulation results demonstrate its advantages compared to a recent VAD algorithm.

In Chapter 5, we conclude this thesis and discuss further research directions. We review the original contribution of this research thesis to the fields of beamforming and VAD and elaborate on possible future directions to this research.
Chapter 2

Related Work and Theoretical Background

In this chapter, we review contemporary beamforming and VAD algorithms, as well as providing a brief theoretical background on relevant methods and metrics in these fields. We start by reviewing the physical theory and models of propagating wave fields and plane waves in particular in Section 2.1. In Section 2.2, we introduce the concept of microphone arrays, presenting their functionality and design considerations. We then present the sampling process of the incoming wave signals by microphone arrays to form time-space signals in Section 2.3. Following the time-space signals, in Section 2.4, we elaborate on microphone array performance measures that are utilized to evaluate microphone array capabilities. In Section 2.5, we review more advanced microphone array design methods that are suitable for wideband signals. Specifically, we review FIR filter based array design methods. In Section 2.6, we present the concept of constant beamwidth beamformers and review two design methods in the field. Achieving a constant beamwidth beamformer is the main goal of Chapter 3. Section 2.7 introduces the background and theory of VAD algorithms as well as obstacles in its implementation. In Section 2.8, we elaborate on one of the major obstacles in VAD, transient noises. Section 2.9 introduces spectral clustering, a common method for data analysis which is the method used for the proposed VAD algorithm in Chapter 4. Finally, in Section 2.10, we elaborate on MFCCs, that are the representation of the sound spectrum in the proposed method in Chapter 4.
2.1 Propagating Wave Fields

In array signal processing, propagating waves carry signals from the source to the array. Therefore, these signals are represented in the time-space domain. The space domain is represented by either the three dimensional Cartesian coordinates \((x, y, z)\) or the three dimensional spherical coordinates \((r, \phi, \theta)\), where \(0 \leq \phi \leq 2\pi\), \(0 \leq \theta \leq \pi\) are the azimuth and elevation angles, respectively. The time domain is represented by \(t\).

Finally, let \(f(t, r)\) denote the time-space representation of the input signal, where \(r\) is the radius-vector in the three-dimensional system. The relations between the Cartesian and spherical coordinates are given in Fig. 2.1.

\[f(t, x, y, z)\] describes the signal impinging the microphone array. The physics of propagation within a homogeneous (constant speed of propagation), dispersion free (no signal degradation due to other frequencies) and lossless (no signal gain attenuation) medium is described by the wave equation:

\[
\nabla^2 f(t, x, y, z) = \frac{1}{c^2} \frac{\partial^2 f(t, x, y, z)}{\partial t^2},
\]

(2.1)

where \(\nabla^2\) represents the Laplacian operator and \(c = \frac{340\pi}{s}\) represents the propagation speed of the wave.
CHAPTER 2. RELATED WORK AND THEORETICAL BACKGROUND

Let \( f_p(t, x, y, z) \) denote a possible solution to (2.1) of a complex exponential form:

\[
f_p(t, x, y, z) = A \exp \left( j(\omega t - k_xx - k_yy - k_zz) \right),
\]

(2.2)

where \( A \) is a complex constant, \( \omega \) denotes a constant temporal frequency and \( k_x, k_y, k_z \) are real constants. Substituting (2.2) into (2.1) yields:

\[
k_x^2 + k_y^2 + k_z^2 = \frac{\omega^2}{c^2}
\]

(2.3)

This solution is called a monochromatic plane wave. The term plane wave is derived from the fact that at any given time, \( t_0 \), the value of \( f(t_0, x, y, z) \) is the same at all points on a plane given by \( k_xx + k_yy + k_zz = \text{const} \). Finally, writing the solution to (2.1) using vector notation we obtain:

\[
f(t, x) = A \exp \left( j(\omega t - kx) \right)
\]

(2.4)

Let \( kx = \text{constant} \) denote planes of constant phase. If the signal \( f(t, x) \) is indeed propagating, planes of constant phase move by small steps of \( \delta x \), with small time increments of \( \delta t \), i.e., \( f(t + \delta t, x + \delta x) = f(t, x) \), which yields:

\[
\omega \delta t - k \delta x = 0
\]

(2.5)

Assuming \( \delta x \) and \( k \) have the same direction, \( k \delta x = |k| |\delta x| \) and \( \frac{|\delta x|}{\delta t} = \frac{\omega}{|k|} \), where \( \frac{|\delta x|}{\delta t} \) can designate the propagation speed of the plane wave. Since \( k \) and \( \omega \) are related by \( |k|^2 = \frac{\omega^2}{c^2} \), we have:

\[
\frac{|\delta x|}{\delta t} = c,
\]

(2.6)

Assuming that \( c > 0 \). Thus, the speed of propagation is indeed \( c \).

Let \( \lambda \) denote the distance the plane wave propagated during a single temporal period of \( T = \frac{2\pi}{\omega} \). This distance is referred to as the wavelength. Due to its spatial traits, after one period the monochromatic plane wave appears the same as one period before, but moved forward.

Let \( k \) denote the wavenumber vector. Its magnitude \( |k| \) expresses the number of cycles in radians per meter of length that the plane wave has exhibited in the propagation direction. Using (2.5) with \( \delta t = \frac{2\pi}{\omega} \), we obtain:

\[
\lambda = \delta x = \frac{2\pi}{|k|}
\]

(2.7)
Therefore, the wavenumber vector can be considered to represent spatial frequency, similarly to the manner \( \omega \) represent temporal frequency.

### 2.2 Microphone Arrays

Microphone arrays consist of sets of microphones positioned in a way to function as a directional acoustic antenna. Microphone arrays are utilized for filtering signals in a space-time field. Filtering is enabled by exploitation of the incoming signals spatial characteristics. A desirable spatial filtering, i.e., beamforming, should result in enhancement of signals of interest, originated in a specific direction, while forcing suppression of undesired signals originated in other directions. Microphone arrays are utilized for solving many signal processing problems: dereverberation, localization of a single source, noise reduction and source separation [31,34,35].

Numerous factors must be taken into consideration when designing a microphone array configuration. Initially, the geometry of the microphone array plays an important role in the formulation of the processing algorithms, as it forces fundamental constraints on the array’s operation. In most cases, the array geometry is the first consideration in array design due to practical and physical constraints of the design. Therefore, the degree of freedom in choosing the array geometry is limited. Nevertheless, in some other crucial problems such as noise reduction or source separation, the geometry of the array may have little importance. For instance, Uniform Linear Arrays (ULA) can only handle one source direction at a time resulting in uncertainties and possibly direction ambiguities. Therefore, ULA configurations must be ruled out when handling several signal sources. Other array geometries such as non-uniform linear arrays and circular arrays, have been studied in the field [36,37]. This work is based on ULA and therefore we elaborate on this geometry.

A conventional beamformer [37] is composed of a ULA of microphones. That is, the microphones are positioned on an axis with a uniform spacing between the microphones. The general expression of the microphone positioning in a ULA is given by:

\[
x_m = m \cdot d, \quad m = 0, 1, \ldots, M - 1
\]

where \( m \) denotes the microphone index; \( x_m \) denotes the position of the microphone having
index \( m \), where \( x_m = 0 \) corresponds to the left-hand side of the array, i.e., the location where the microphone indexed by \( m = 0 \) is positioned; \( d \) denotes the spacing between microphones, and \( M \) is the array size, i.e., the number of microphones in the array, as demonstrated in Fig. 2.2.

\[
x = 0 \quad \ldots \quad (M - 1) d
\]

\[
\begin{array}{c}
\bigcirc \\
m = 0 \\
\end{array} \quad d \quad \begin{array}{c}
\bigcirc \\
M - 1
\end{array}
\]

Figure 2.2: A uniform linear array of size \( M \) and spacing \( d \).

### 2.3 Time-Space Signals

The incoming signals are spatially sampled by the array microphones, then the samples are processed to attenuate signals from undesired directions and extract the signal from a desired direction. A spatial response of the microphone array is obtained via a beam (main-lobe) directed to the desired signal while nulls are directed to the undesired signals. Figure 2.3. illustrates a beamformer design based on a ULA of Fig. 2.2.

Let \( f(t, x) \) denote the set of signals sampled by the microphone array at time \( t \), expressed by

\[
f(t, x) = [f(t, x_0), \ldots, f(t, x_{M-1})]^T,
\]

where \( x_m \) denotes the microphone position, \( t \) denotes the continuous-time variable, and \( T \) denotes the transpose operation. The beamformer output \( y(t) \) is expressed by:

\[
y(t) = \sum_{m=0}^{M-1} f(t, x_m) w_m^*,
\]

where \( w_m \) is a complex weight of the microphone having index \( m \), as shown in Fig. 2., and \( * \) denotes complex conjugation. The simplest beamformer design architecture has uniform weights, \( w_m = \frac{1}{M}, \ m = 0, \ldots, M - 1 \).
Let $f_{\omega}(t, x) = e^{j\omega t}$ denote a plane wave propagating at angular frequency $\omega$, and let $\theta, \theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, denote the direction of arrival (DOA) angle measured with respect to the broadside of the linear array, as shown in Fig. 2. The wave signals spatially sampled by the microphone array inputs are given by:

$$f_{\omega}(t, x) = [f_{\omega}(t - \tau_0), \ldots, f_{\omega}(t - \tau_{M-1})]^T, \quad \tau_m = \frac{\sin(\theta) \cdot x_m}{c}$$

(2.11)

where $\tau_m$ is the propagation delay for the incoming signal and $c = 340 \frac{m}{s}$ denotes the velocity of sound propagation in air. A value of $\tau_m = 0, \forall m$ implies a DOA of $\theta = 0$, i.e., a plane wave parallel to the array, propagating perpendicularly to the array. Let $\kappa = \frac{\omega}{c} \sin(\theta) = \frac{2\pi}{\lambda} \sin(\theta)$ denote the wavenumber for plane waves in a locally homogeneous medium, where $\lambda$ denotes the wavelength corresponding to the angular frequency $\omega$, and let $\mathbf{v}(\kappa)$ denote the array manifold vector [37], featuring all of the spatial characteristics of the microphone array. Based on (3.4), and the definition of $\kappa$ above, the manifold vector can be expressed as:

$$\mathbf{v}(\kappa) = [e^{-j\kappa x_0}, \ldots, e^{-j\kappa x_{M-1}}]^T$$

(2.12)
2.4 Microphone Array Performance Measures

Numerous performance measures are utilized for evaluating the microphone array capabilities. Each of the measures aims to quantify a significant aspect of either the response of an array to the signal environment or of the sensitivity to an array design error.

First, we review the concept of beampatterns. Beampatterns are the main tool in assessing an array performance as they express the beamformer response for various frequencies and signal angles. Let $P(\omega, \theta)$ denote the frequency and DOA dependent beamformer response. It is given by:

$$
P (\omega, \theta) = \sum_{m=0}^{M-1} e^{-j\omega \tau_m} w_m^* = \mathbf{w}^H \mathbf{v} (\omega, \theta),
$$

(2.13)

where $\mathbf{w} = [w_0, \ldots, w_{M-1}]^T$ and $^H$ denotes Hermitian transpose operation, and let $D (\omega, \theta)$ denote the beampattern of a given beamformer. It is expressed by:

$$
D (\omega, \theta) = 20 \log_{10} |P (\omega, \theta)|,
$$

(2.14)

where $|\cdot|$ denotes the absolute value. For a given angular frequency $\omega$, the beampattern $D (\omega, \theta)$ is a function of the angle $\theta$ and the beamwidth is measured in terms of $\theta$. In this work, the beamwidth is measured between the two lowest values at both sides of the main lobe.

A major challenge in practical beamformer applications is the potential sensitivity to mismatches between the actual array attributes and the model used to derive the desired beamformer. In practical applications, mismatches can occur either by array location perturbations, production faults or filter perturbations. The sensitivity function often used as a criterion for assessing the affect of mismatches on the array response is defined in [37] by:

$$
T_{se} = A_w^{-1} = \| \mathbf{w} \|^2,
$$

(2.15)

where $A_w^{-1}$ is the inverse expression of the white noise gain given by $A_w = \text{SNR}_{\text{out}}(k) / \text{SNR}_{\text{in}}(k)$ and $\mathbf{w}$ is the weight vector corresponding to all of the FIR filter channels in the $k$-th frequency bin. Therefore, as the white noise gain increases, the sensitivity decreases and the array would be more robust to mismatch.
2.5 Finite Impulse Response Filter Based Filtering

Delay-and-sum beamformers, as shown in Fig. 2.3., utilize a single weight for each microphone. This makes them ineffective when dealing with wideband signals, the signals of interest in speech and audio processing, as each of the signals is composed of various frequency components. In order to design a beamformer for wideband signals, the weight values in (3.6) must be altered for different frequencies in order to obtain the desired beamformer output. That is, the weights should be frequency dependent, i.e., in the form of \( w(\omega) = [w_0(\omega), \ldots, w_{M-1}(\omega)]^T \). This can be achieved by introducing discrete FIR filters [19,24].

FIR filters perform temporal filtering in order to compensate for the phase differences of the input wideband signals’ various frequency components. Fig. 3. shows the realization of frequency dependent weights by FIR filters connected to the microphone array.

Let \( y(t) \) denote the output of the FIR-based beamformer. It is expressed by:

\[
y(t) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(t - nT_s, x_m) \cdot w^*_m,n,
\]

where \( N \) denotes the number of FIR filter coefficients connected to each of the \( M \) microphones, \( w^*_m,n \) is the \( n \)-th coefficient of the FIR filter connected to the microphone indexed by \( m \), and \( T_s \) denotes the delay between adjacent filter elements. Given a complex plane wave signal, the beamformer output is given by:

\[
y(t) = e^{j\omega t} P(\omega, \theta) = e^{j\omega t} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} e^{-j\omega(\tau_m + nT_s)} \cdot w^*_m,n,
\]

Let \( v_m(\omega, \theta) \) denote a stacked array manifold vector of dimension \( M \cdot N \), where each subvector of dimension \( M \) represents the array manifold vector associated with a specific FIR filter coefficient in (3.10), i.e., the first subvector, \( [e^{-j\omega(\tau_0)}, \ldots, e^{-j\omega(\tau_{M-1})}]^T \), is associated with the coefficient indexed by \( n = 0 \) and all of the array microphones, indexed by \( m = 0, \ldots, M - 1 \). Thus, it is denoted by \( v_0(\omega, \theta) \). The second subvector, \( [e^{-j\omega(\tau_0+T_s)}, \ldots, e^{-j\omega(\tau_{M-1}+T_s)}]^T \), is associated with the coefficient indexed by \( n = 1 \), and all of the array microphones. Thus, it is denoted by \( v_1(\omega, \theta) \), and so on. This form of
expression is called vector stacking, and \( \mathbf{v}_s(\omega, \theta) \) is given by:

\[
\mathbf{v}_s(\omega, \theta) = \begin{bmatrix} 
\mathbf{v}_0(\omega, \theta) \\
\mathbf{v}_1(\omega, \theta) \\
\vdots \\
\mathbf{v}_{N-1}(\omega, \theta) 
\end{bmatrix}.
\] (2.18)

Given a ULA of \( M \) microphones, microphone spacing \( d \), and \( M \) FIR filters, each composed of \( N \) coefficients, and connected to a respective microphone. Then, from (3.10) and (3.4), the beamformer response can be expressed as:

\[
P(\omega, \theta) = \sum_{m=0}^{M-1} e^{-j\omega \tau_m} \sum_{n=0}^{N-1} e^{jn\omega T_s} \cdot w_{m,n}^* = \mathbf{w}_s^H \mathbf{v}_s(\omega, \theta),
\] (2.19)

where \( \mathbf{w}_s \) denotes the composite stacked weight vector of dimension \( M \cdot N \) created by vector stacking, having \( \mathbf{w}_0 = [w_{0,0}, w_{1,0}, \ldots, w_{M-1,0}]^T \) as its first subvector, \( \mathbf{w}_1 = [w_{0,1}, w_{1,1}, \ldots, w_{M-1,1}]^T \) as its second subvector, and so on. The design specification of the FIR filters will be addressed in Chapter 3. Figure 3. illustrates an FIR beamformer architecture.

Figure 2.4: FIR beamformer architecture, where D denotes a delay element.
CHAPTER 2. RELATED WORK AND THEORETICAL BACKGROUND

The second summation in (3.12) can be expressed in terms of the Discrete-time Fourier transform (DTFT) of the $m$-th channel filter coefficients as follows:

$$\sum_{n=0}^{N-1} e^{j\omega T_s} w_{m,n}^* = W_m^* (e^{j\omega T_s})^*, \quad (2.20)$$

where $W_m (e^{j\omega T_s})$ denotes the DTFT of $w_{m,n}$ with the summation variable $n$. Hence, the expression for $P(\omega, \theta)$ in (3.12) can be written as:

$$P(\omega, \theta) = \sum_{m=0}^{M-1} e^{-j\omega \tau_m} \cdot W_m^* (e^{j\omega T_s}). \quad (2.21)$$

Since the channel filters have each a finite number of coefficients ($N$), their frequency responses can be represented by the Discrete Fourier transform (DFT) of size $N$ of each filter coefficients. This corresponds to sampling $\omega$ at $\frac{2\pi}{N T_s} k, \; k = 0, \ldots, N-1$. Examining the beamformer response for frequencies corresponding to the bin centers defined by the DFT, we obtain:

$$P(k, \theta) = \sum_{m=0}^{M-1} e^{-j\frac{2\pi \tau_m}{N T_s} k} \cdot W_m^* (e^{j\frac{2\pi k}{N}}), \; k = 0, \ldots, N-1, \quad (2.22)$$

where $k$ denotes the bin index of the DFT. This expression will be the basis for controlling the beamformer frequency response by specifying the frequency response of the FIR channel filters.

2.6 Constant Beamwidth Beamformers

Constant beamwidth beamformers are beamformers that have a constant beamwidth over a range of frequencies in a wideband. The beamwidth of a ULA decreases as frequency increases. Thus, in narrowband applications, the inverse proportionality of the beamwidth with respect to frequency is generally insignificant. A constant beamwidth is desirable in wideband applications such as speech processing with microphone arrays and multiple user conferencing. Traditional beamformer design approaches did not provide sufficient performance in terms of beamwidth constancy and other performance criteria. Therefore, numerous contemporary approaches to robust beamformer designs were proposed. The minimum variance distortionless response (MVDR) [14,15], linearly constrained minimum
variance (LCMV) [16] and least-squares [17, 18] are several widely used beamforming design techniques developed for improving the traditional methods.

One of the popular beamformer design approaches is the FIR based beamformer [19–23]. FIR beamformers perform temporal filtering that results in a frequency dependent response. The filtering is performed in order to compensate for the phase difference of various frequency components. It has been shown [24,25] that utilizing narrowband FIR filters in beamformer designs can produce robust beamformers.

### 2.6.1 Multi Beamforming with FIR Filtering Approach

A beamformer can be modeled as a tapped delay line, where the delays determine the beam steering angle. Arrays intended for wideband uses are typically capable of beam steering, i.e., the main-lobe can be rotated from the broadside for different delay values. In conventional delay lines, the delays are constant as a function of frequency, resulting in a frequency independent steering angle. Modeling the beamformer as a tapped all-pass filter line, results in frequency dependent delays. These delays are equivalent to the phase delays of the constituent all-pass filters. If these all-pass filters have linear phase with zero phase at DC, the delays are constant, implying that the fixed delay line is a specific case of this revised beamformer model.

In multi-beamforming, a broadside beam and several steered beams are simultaneously formed, where the steered beams are formed symmetrically with respect to broadside at fixed steering angles. The inter-null beamwidth of the resultant multi-beam is approximately given by the angular distance between the outermost nulls of the outermost beams. This is only an approximation due to the effects of interfering side-lobes from the constituent beams. Since the outermost beams narrow as frequency increases, the beamwidth of the multibeam decreases as well, by roughly the same amount as that of a single beam. Thus, this technique does not yield beamwidth constancy.

A constant beamwidth can be obtained by introducing low-pass filters to the microphone array. The cutoff frequency of these filters depend on the position of their corresponding element with respect to the array center, i.e., the farther the element is from the center, the lower the cutoff of its filter. At low frequencies, all of the elements are used and for increasing frequency, the array shortens in length since the signals corresponding
to the outer elements are progressively attenuated. The design of the elemental low-pass filters has been neglected in the literature due to the apparent computational difficulty.

In the constant beamwidth beamformer design method presented in [38], it is shown that constant beamwidth beamforming can be obtained by incorporating the capability of frequency-dependent beam steering into the multi-beamforming approach. The multi-beambeamwidth decreases as frequency increases due to the narrowing of the outermost beams. If the steering angles are made frequency-dependent, the narrowing can be prevented by steering the outermost beams to larger angles as frequency increases. In this way the angular position of the outermost nulls can be kept constant, resulting in a constant beamwidth.

In the revised beamformer, a constant beamwidth can be achieved over an extended frequency range. At the low end of the frequency band, $2r + 1$ beams are formed simultaneously at broadside. As frequency increases, the outermost beams are steered to keep the outermost nulls stationary. The inner beams are steered proportionally to maintain the integrity of the beam shape.

For a uniformly excited array, the approximate null-to-null beamwidth of the multibeam is twice the angle at which the outermost null is located. The beamwidth is given by:

$$\theta_{BW} = 2 \sin^{-1} \left( \frac{2\pi}{M\omega\tau_0} + \frac{\tau(\omega)}{\tau_0} \right), \quad (2.23)$$

where $\tau(\omega)$ denotes the delay of the all-pass filters. Let $\phi_r$ denote the frequency-dependent steering angle of the outermost beam expressed by:

$$\phi_r = 2 \sin^{-1} \left( \frac{2\pi}{M\omega\tau_0} + \frac{\tau(\omega)}{\tau_0} \right) \quad (2.24)$$

$\tau(\omega)$ is chosen to keep the arcsine argument in (2.24) constant as a function of frequency. Choosing such a $\tau(\omega)$ results in a constant beamwidth beamformer.

Let $\Phi(\omega)$ denote the phase function corresponding to $\tau(\omega)$ expressed by:

$$\Phi(\omega) = \frac{2\pi}{M} \left( 1 - \frac{\omega}{\omega_0} \right), \quad (2.25)$$

where $\omega_0$ denotes the lower frequency limit, i.e., the frequency at which all of the constituent beams are formed at broadside. This phase function is affine, i.e., linear in $\omega$ but have a DC offset.
Since $2r + 1$ beams are being formed simultaneously, the far field response of the multi-beamformer is given by:

$$A_m(\omega, \theta) = \sum_{i=-r}^{r} \sum_{n=-N}^{N} a_n \exp(-jn\omega \tau_0 \sin(\theta)) \exp\left(\frac{jn\Phi(\omega) i}{r}\right)$$  \hspace{1cm} (2.26)

where $\Phi(\omega)$ denotes the proper phase function to achieve constant beamwidth given the taps $a_n$. Interchanging the summation order and invoking the symmetry of the multi-beam yields:

$$A_m(\omega, \theta) = \sum_{n=-N}^{N} B_n(\omega) \exp(-jn\omega \tau_0 \sin(\theta))$$

$$B_n(\omega, \theta) = a_n \left[1 + \sum_{i=-r}^{r} 2 \cos\left(\frac{n\Phi(\omega) i}{r}\right)\right]$$  \hspace{1cm} (2.27)

The zero-phase lowpass filters $B_n(\omega)$ are the appropriate elemental lowpass filters discussed in Section 3.2. If $B_n(\omega)$ are used as frequency dependent shading weights for an ULA, a nearly constant inter-null beamwidth results.

The use of the elemental lowpass filters results in a low-pass broadside response, which is undesirable since it reduces the fidelity of signals corresponding to the main beam. This nonuniformity can be removed by a compensation filter given by:

$$G(\omega) = \frac{1}{\sum_{n=-N}^{N} B_n(\omega)}$$  \hspace{1cm} (2.28)

### 2.6.2 Least Squares Approach

The least squares (LS) approach is a prominent design method for both the FIR filters and broadband beamformers [18]. It can provide a closed-form solution to the problem and is more computationally efficient compared with the convex optimization method. Nevertheless, most contemporary LS methods are utilized for the design of general broadband beamformers and not necessarily for designing constant beamwidth beamformers.

In the work presented in [18], a spatial variation (SV) formulation is introduced into the cost function as a frequency invariance controlling element. As a result, the unity response used as the desired response over the whole signal bandwidth is replaced by constraining the response at a single reference frequency. Then, a linearly constrained least squares formulation is derived and it can be solved either by the Lagrange multipliers method, which is a direct solution to the LS problem with linear constraints, or by transforming
the constrained optimization problem into an unconstrained problem by decomposing the coefficient vector into two orthogonal components.

Let \( J_{LS} \) denote a LS cost function given by:

\[
J_{LS} = \int_{\Omega_i} \int_{\Theta} F(\Omega, \theta) \left| w^T s(\Omega, \theta) - D(\Omega, \theta) \right|^2 d\Omega d\theta \tag{2.29}
\]

where \( \Omega_i \) and \( \Theta \) represent the frequency and angle range of interest, respectively, \( F(\Omega, \theta) \) denotes a positive real-valued weighting function and \( D(\Omega, \theta) \) denotes a desired response function.

Let \( (\Omega_n, \theta_k) \) denote the grid uniformly chosen from the continuous frequency and angle ranges so (2.29) can now be expressed by:

\[
J_{LSD} = \sum_{\Omega_n \in \Omega_i} \sum_{\theta_k \in \Theta_m} \left| w^T s(\Omega_n, \theta_k) - 1 \right|^2 + \alpha \sum_{\Omega_n \in \Omega_i} \sum_{\theta_k \in \Theta_s} \left| w^T s(\Omega_n, \theta_k) \right|^2 \tag{2.30}
\]

where \( \Theta_s \) denotes the side-lobe region and \( \Theta_m \) d denotes the main-lobe region. (2.30) can be expressed as a quadratic function:

\[
J_{LSD} = w^T Q_{LS} w - 2w^T a + d_{LS}, \tag{2.31}
\]

with

\[
Q_{LS} = \sum_{\Omega_n \in \Omega_i} \sum_{\theta_k \in \Theta_m} S_R(\Omega_n, \theta_k) + \alpha \sum_{\Omega_n \in \Omega_i} \sum_{\theta_k \in \Theta_s} S_R(\Omega_n, \theta_k), \tag{2.32}
\]

\[
S_R(\Omega_n, \theta_k) = \text{Re} \left\{ S_R(\Omega_n, \theta_k) = s(\Omega_n, \theta_k) s(\Omega_n, \theta_k)^H \right\} \tag{2.33}
\]

\[
a = \sum_{\Omega_n \in \Omega_i} \sum_{\theta_k \in \Theta_m} s_R(\Omega_n, \theta_k), \tag{2.34}
\]

\[
d_{LS} = \sum_{\Omega_n \in \Omega_i} \sum_{\theta_k \in \Theta_m} 1 \tag{2.34}
\]

Finally, the minimization solution of (2.30) is given by:

\[
w_{LS} = Q_{LS}^{-1} a \tag{2.35}
\]

Here, there is no constraint that guarantees a beamwidth constancy. Thus, the introduction of a frequency invariance controlling element into the design is necessary and it is denoted as \( SV \) and defined as:

\[
SV = \sum_{\Omega_n \in \Omega_i} \sum_{\theta_k \in \Theta_{FI}} \left| w^T s(\Omega_n, \theta_k) - w^T s(\Omega_r, \theta_k) \right|^2 \tag{2.36}
\]
where $\Theta_{FI}$ denotes the direction range in which beamwidth constancy is considered and $r$ is a fixed reference frequency. The parameter $SV$ is a measurement of the Euclidean distance between the response at the fixed reference frequency $r$ and that at all the other operating frequencies over a range of directions in which beamwidth constancy is considered. When the beamformer has a constant beamwidth, the value of $SV$ will be zero.

Since the beamwidth constancy property is expected to be held also in the side-lobe region, there is only a need to minimize the spectrum energy of the beamformer at the reference frequency $r$ over the side-lobe region, which is given by:

$$ J_1 = \sum_{\theta_k \in \Theta_s} |w^T_s(\Omega_r, \theta_k)|^2 $$ \hspace{0.5cm} (2.37)

Moreover, the unity response over the whole frequency band of interest in the look direction in the original formulation can be replaced by constraining the response of the beamformer at the reference frequency in the look direction $\theta_r$ to be unity, which is given by:

$$ w^T_s(\Omega_r, \theta_r) = 1 $$ \hspace{0.5cm} (2.38)

Then, a constrained LS formulation of the constant beamwidth beamformer design problem is obtained by combining (2.36), (2.37) and (2.38):

$$ J_{CLS} = \sum_{n=0}^{N-1} \sum_{k=0}^{K-1} |w^T_s(\Omega_n, \theta_k) - w^T_s(\Omega_r, \theta_k)|^2 + \beta \sum_{\theta_k \in \Theta_s} |w^T_s(\Omega_r, \theta_k)|^2 $$ \hspace{0.5cm} (2.39)

s.t $w^T_s(\Omega_r, \theta_r = 1)$

where $N$ and $K$ denote the number of samples chosen uniformly over the frequency and the angle ranges considered for beamwidth constancy, respectively, and $\beta$ is a trade off parameter between the beamwidth constancy property and the sidelobe attenuation. (2.39) can be written as:

$$ J_{CLS} = w^T Q_{CLS} w $$ \hspace{0.5cm} s.t $s(\Omega_r, \theta_r = 1)^H w = 1 $$ \hspace{0.5cm} (2.40)

where,

$$ Q_{CLS} = \sum_{n=0}^{N-1} \sum_{k=0}^{K-1} \text{Re} \left\{ (s(\Omega_n, \theta_k) - s(\Omega_r, \theta_k)) (s(\Omega_n, \theta_k) - s(\Omega_r, \theta_k))^H \right\} + \beta \sum_{\theta_k \in \Theta_s} S(\Omega_r, \theta_k) $$ \hspace{0.5cm} (2.41)
Since \( s(\Omega_r, \theta_r) \) is complex valued, the single complex constraint can be changed into two real constraints:

\[
C^T w = f
\]

\[
C = [s(\Omega_r, \theta_r)_R, s(\Omega_r, \theta_r)_I], \quad f = [1, 0]^T
\]

And (2.40) can be rewritten as:

\[
J_{CLS} = w^T Q_{CLS} w \\
s.t. \quad C^T w = f
\]

which can be solved by the Lagrange multipliers method:

\[
w = Q_{CLS}^{-1} C \left( C^T Q_{CLS} C \right)^{-1} f
\]

Additionally, the constrained optimization problem in (2.43) can be transformed into an unconstrained problem by decomposing the coefficient vector \( w \) into two orthogonal components, \( w_q \) and \( -v \) as:

\[
w = w_q - v,
\]

where \( w_q \) lies in the range of matrix \( C \) and \( v \) is in the null space of \( C \), i.e. the space of all \( v \) fulfilling \( C^T v = 0 \). The range and null space of a matrix span the entire space. Thus, this decomposition can be used to represent any \( w \). \( C^T w_q = f \) is a condition that must be met in order to meet the constraint equation (2.42) and have:

\[
w_q = (C^T)\dagger f = C (C^T C)^{-1} f,
\]

where \( \dagger \) denotes the pseudo-inverse operator.

\( v \) can be expressed as a linear combination of the basis vectors of the null space of \( C \):

\[
v = Bw_a
\]

where \( B \) satisfying \( C^T B = 0 \) can be obtained from \( C \) using orthogonalization methods, such as the singular value decomposition, and \( w_a \) is given by:

\[
w_a = (B^T Q_{CLS} B)^{-1} B^T Q_{CLS} w_q
\]

Finally, the solution for \( w \) can be obtained with \( w_a, w_q \) and \( B \).
2.7 Voice Activity Detection

Time Assignment Speech Interpolation (TASI) [39] was the first method to use voice activity detection for enhancing trans-Atlantic telecommunication capacity by switching users onto idle channels. Today, the most widely used cases of VAD are the Global System for Mobile Communication (GSM) [7] for mobile communication and the G.729 [8], used mainly in Voice over IP. GSM includes two VAD operations: First, computation of the Signal-to-Noise Ratio (SNR) in nine bands and applying a threshold to these values. Second, calculation of various parameters such as noise and channel power and voice metrics. The algorithm then thresholds the voice metrics using a varying threshold which varies according to the estimated SNR. G.729 utilizes basic features such as the energy of the signal and zero-crossing rate. These standards demonstrate fairly good performance, especially for clean signals and are therefore widely used in communication. Nevertheless, their performances degrade in presence of environmental noise, even for relatively high SNR values.

To overcome this shortcoming, several statistical model based VAD algorithms have been proposed in the last two decades [26, 27, 40, 41]. Although much progress has been made in improving VAD algorithms performance in the presence of environmental noise, overcoming transient noise still remains a big obstacle.

Let $x_s(n)$ denote a speech signal and $x_n(n)$ denote an additive interfering background noise signal. The microphone input signal, denoted by $y(n)$ is thus given by a signal contaminated with additive background noise, given by:

$$y(n) = x_{sp}(n) + x_n(n)$$  \hspace{1cm} (2.49)

The signal is processed using consecutive frames in the STFT domain, such that (2.49) is given in by:

$$Y(t,k) = Y_s(t,k) + Y_n(t,k), \quad k = (1, 2, ..., K_m), \quad t = (1, 2, ..., N)$$  \hspace{1cm} (2.50)

where $Y(t,k)$, $Y_s(t,k)$ and $Y_n(t,k)$ denote the STFT coefficients of $y(n)$, $x_{sp}(n)$ and $x_n(n)$, respectively, $k$ denotes the frequency bin index, and $t$ the time frame index. Let $\mathcal{H}_0$ and $\mathcal{H}_1$ denote speech absence and presence hypotheses, respectively, and let $f(\cdot; \mathcal{H}_0)$ and
Let $f(\cdot; \mathcal{H}_1)$ be PDF of the input signal in the STFT domain conditioned on the hypotheses, defining a log of the likelihood ratio:

$$\Lambda_t(k) = \log \left( \frac{f(Y(t,k); \mathcal{H}_1)}{f(Y(t,k); \mathcal{H}_0)} \right). \quad (2.51)$$

In the method presented in [26], both signal noise are assumed to be complex Gaussian uncorrelated processes in the STFT domain expressing the PDF functions by:

$$f(Y(t,k); \mathcal{H}_0) = \frac{1}{\pi \lambda_n(t,k)} \exp \left( -\frac{|Y(t,k)|^2}{\lambda_n(t,k)} \right), \quad (2.52)$$

for the speech absence hypotheses, and:

$$f(Y(t,k); \mathcal{H}_1) = \frac{1}{\pi [\lambda_s(t,k) + \lambda_n(t,k)]} \exp \left\{ -\frac{|Y(t,k)|^2}{\lambda_s(t,k) + \lambda_n(t,k)} \right\}, \quad (2.53)$$

for the speech presence hypotheses, where $\lambda_s(t,k)$ and $\lambda_n(t,k)$ denote the spectral variance of speech and noise, respectively. For the Gaussian model, $\Lambda_t(k)$ is given in [26] by:

$$\Lambda_t(k) = \left( \gamma(t,k) \cdot \xi(t,k) \right) - \log (1 + \xi(t,k)), \quad (2.54)$$

where $\xi(t,k) \equiv \frac{\lambda_s(t,k)}{\lambda_n(t,k)}$ and $\gamma(t,k) \equiv \frac{|\lambda_s(t,k)|^2}{\lambda_n(t,k)}$ are the a-priori and a-posteriori SNR, respectively.

Let $\lambda_n(t,k)$ denote the estimated spectral variance, and let $Y_s(t,k)$ denote the Minimum mean square error (MMSE) estimator of the speech signal, the estimated a-priori SNR is given by:

$$\xi(t,k) = \alpha \frac{|Y_s(t-1,k)|^2}{\lambda_n(t-1,k)} + (1-\alpha)P[\gamma(t,k) - 1], \quad (2.55)$$

where $\alpha \in (0,1)$ is a system smoothing parameter, $\gamma(t,k) - 1$ is the maximum likelihood estimator of $\xi(t,k)$, and $P(x)$ is utilized for keeping non-negative values of the estimator [42].

The estimated log of the likelihood ratio in (2.54) is then averaged over the frequency bins, and the arithmetic mean of the log of the likelihood ratio is given by:

$$\Lambda_t = \frac{1}{K_f} \sum_{k=1}^{K_f} \Lambda_t(k), \quad (2.56)$$

where $\Lambda_t$ is a feature indicating voice activity in a frame indexed by $t$. In [26], voice activity is detected by comparing $\Lambda_t$ to a threshold such that speech is present in the frame if $\Lambda_t$ is greater than a certain threshold. The decision is then smoothed using a Hang-Over scheme which assumes a strong correlation in the occurrences of consecutive speech frames [26].
2.8 Transient Noises

Transient noise signals are high energy, non-stationary signals, such as coughing, door knocking and typing. These signals are characterized by short outbreaks of sound that vary faster than speech [32]. Traditional methods reviewed in the previous section, assume that noise signals vary slowly with respect to speech signals. Thus, transients are often detected as speech signals by these algorithms. Traditional algorithms utilize spectral variance estimation of the noise, $\lambda_n(t,k)$, and their performance confide in the accuracy of noise statistics estimations. Moreover, for low SNR values with non-stationary noise, speech components may have substantial noise in both speech segments, in the form of distorted speech and noise segments, in the form of non-existent speech features. These degrade the traditional VAD algorithm performance and require the use VAD algorithms to distinguish speech absence from speech presence.

One of the prominent methods for the estimation of the spectral variance of the noise is the Minima Controlled Recursive Averaging (MCRA) method [28,29], where the spectral variance is recursively smoothed in speech absence intervals. Although MCRA has been successful in estimation of the spectral variance of various quasi-stationary noises, it ineffective for the estimation of spectral variance of transients.

Spectral variance estimation of the noise is given by smoothing the energy of the noisy signal over time in speech absent intervals, while keeping previous values of the estimate in speech intervals:

$$
\hat{\lambda}_n(t+1,k) = \begin{cases} 
\alpha_n\hat{\lambda}_n(t,k) + (1 - \alpha_n)|Y(t,k)|^2 ; & H'_0 \\
\hat{\lambda}_n(t,k) ; & H'_1 
\end{cases}
$$

(2.57)

where $\alpha_n \in (0,1)$ is a smoothing parameter, $H'_0$ and $H'_1$ are speech absence and speech presence hypotheses for noise spectrum estimation, respectively. Here, a decision that speech is absent in a frame is taken with a lower confidence, i.e., $f(H'_0|Y(t,k)) \leq f(H_0|Y(t,k))$. Both $H'_0$ and $H'_1$ are utilized for avoiding the smoothing of $\hat{\lambda}_n$ over speech components due to wrong speech detections. Since $H'_0$ and $H'_1$ are unknown, from (2.57) we obtain:

$$
\hat{\lambda}_n(t+1,k) = \hat{\lambda}_n(j,k)p'(t,k) + \left[\alpha_n\hat{\lambda}_n(t,k) + (1 - \alpha_n)|Y(t,k)|^2\right] (1 - p'(t,k))
$$

(2.58)
where $p'(t, k)$ denotes the probability of speech presence conditioned on the noisy signal, i.e., $p'(j, k) = f(H'_1|Y(t, k))$.

$p'(t, k)$ is estimated by the noisy signal energy, smoothed in time and frequency.

Let $S_f(t, k)$ be a local energy of the signal smoothed in the frequency domain using a window of length $2\tilde{w} + 1$:

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

where $b(i)$ denotes a window function. Let $S(t, k)$ denote the energy of the noisy signal, which is smoothed in time and frequency, and is expressed by:

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

where $\alpha_s \in (0, 1)$ is a smoothing parameter. Here, the presence of speech is assumed when the smoothed signal energy is greater than the energy of the noise.

Let $I(t, k)$ denote an indicator of speech given by:

$$I(t, k) = \begin{cases} 
1 & ; \quad S(t, k) > \delta S_{\text{min}}(t, k) \\
0 & ; \quad \text{otherwise}
\end{cases},$$

where $\delta$ denotes some constant parameter and $S_{\text{min}}(t, k)$ denotes the minimum of the noisy signal’s smoothed energy.

We remark that the main assumption in the estimation of the spectral variance of the noise lies in equation (2.61), 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 [30, 43]. 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 [32], the arithmetic mean of the log likelihood ratio, $\Lambda_j$, 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.

2.9 Spectral Clustering

Spectral clustering is a common method for data analysis. It utilizes the spectrum, i.e., eigenvalues of the similarity matrix of the data to perform dimensionality reduction prior to clustering the data in less dimensions. The similarity matrix is provided as an input and composed of a quantitative assessment of the relative similarity of each pair of points in the dataset.

Similarity graphs are a common data representation method in signal processing. Let $G = (V, E)$ denote a weighted graph, where $V$ and $E$ represent the set of vertices and edges of the graph, respectively. Each vertex $v_i$ represents a data point $Y(:, i)$ and each edge $e_{ij}$ between two vertices $v_i$ and $v_j$ has a non-negative weight $W(i,j) \geq 0$. These weights are a measure of similarity between points. We assume that the graph $G$ is undirected, i.e., $W(i,j) = W(j,i)$.

Let $W$ denote a similarity matrix whose $(i,j)$-th element equals to $W(i,j)$. In [44], a spectral clustering algorithm using the eigenvectors of the matrix $D^{-1/2}WD^{-1/2}$ was proposed. $D$ denotes a diagonal matrix whose $i$-th diagonal element is $\sum_{j=1}^{N} W(i,j)$. Let $K$ denote the number of desired clusters and $U$ denote a matrix consisting of the first $K$ eigenvectors of $D^{-1/2}WD^{-1/2}$. These $K$ eigenvectors represent the $K$ largest eigenvalues of $D^{-1/2}WD^{-1/2}$. The clustering is implemented by applying a k-means algorithm on either $U$, where each point is represented by a row of $U$, or on $V = D^{1/2}U(U^T DU)^{-1}$, where each point is represented by a row of $V$.

The main part of spectral clustering algorithms is the aforementioned similarity matrix expressed by a Gaussian kernel:

$$W(i,j) = \exp \left( -\frac{\| Y(:,i) - Y(:,j) \|^2}{\sigma} \right), \quad (2.62)$$

where $Y(:,i)$ is the $i$-th data point and $\sigma$ is commonly done manually. Other algorithms [44] introduced a method for estimating the parameters of the kernel based on minimization of a cost function that characterizes how close the eigenstructure of the similarity matrix.
Mel Frequency Cepstrum Coefficients

Mel Frequency Cepstrum Coefficients (MFCCs) are coefficients that form a representation of the short-term power spectrum of a sound. MFCC is based on a linear cosine transformation of a log power spectrum on a non-linear Mel scale frequency, thus convenient for human auditory applications.

Let $y_i(n)$ denote a sampled speech signal, divide into $M$ short frames, $1 < i < M$, and $N$ samples each, and let $Y_i(K)$ denote the complex DFT of $y_i(n)$ expressed by:

$$ Y_i(k) = \sum_{n=1}^{N} y_i(n) w(n) e^{-j\frac{2\pi kn}{N}}, \quad 1 \leq k \leq K, \quad (2.63) $$

where $w(n)$ is an analysis window of size $N$, and $K$ denotes the length of DFT.

Let $P_i(k)$ denote the power spectrum of the $i$-th frame given by:

$$ P_i(k) = \frac{1}{N} |Y_i(k)|^2 \quad (2.64) $$

Next, the Mel-spaced filterbank is calculated. This filterbank is of size $m \times K$, where $m$ denotes the number of MFCCs and $K$ is the length of the DFT. The filterbank is composed of a set of $m$ triangular filters applied to the power spectral estimate in (2.64). To calculate the filterbank energies, each filterbank is multiplied with the power spectrum and then added up to the coefficients. This yields a vector of length $m$, indicating the amount of energy in each filterbank. The log operator is then applied on each of the size $m$ vector values and finally, a DCT is applied, resulting in the desired MFCCs.
Chapter 3

FIR-Based Symmetrical Acoustic Beamformer With a Constant Beamwidth

In this chapter, we present a method for the design of a constant beamwidth beamformer utilizing FIR filters. We start with a relevant mathematical background and formulate the problem in Section 3.1. In Section 3.2, we review our proposed method for an FIR filter based beamformer and elaborate on the design steps. We review the experimental results of the proposed method in Section 3.3 and present the various scenarios we had simulated, as well as compare the proposed method to other available methods in the field. Finally, we conclude the FIR based constant beamwidth beamformer research chapter and discuss future research directions in Section 3.4.

3.1 Background and Problem Formulation

In this section, we provide background material on filter-based beamformer design and formulate the constant beamwidth problem. Microphone arrays are utilized for filtering signals in a space-time field. Filtering is enabled by exploitation of the incoming signals spatial characteristics and may be expressed in terms of angle and wave number dependency. A desirable spatial filtering, i.e., beamforming, should result in enhancement of signals of interest, originated in a specific direction, while forcing suppression of undesired
signals originated in other directions. Many factors must be taken into consideration when designing a beamformer. First of all, the array geometry forces fundamental constraints on the beamformer’s operation. In most cases, the array geometry is the first consideration in array design due to practical and physical constraints. Therefore, the degree of freedom in choosing the array geometry is limited. Secondly, the weights applied to array microphones determine the spatial filtering characteristics of the beamformer as well as the array performance in terms of measures like sensitivity and directivity.

### 3.1.1 Microphone Array Configuration

A conventional beamformer [37] is composed of a Uniform Linear Array (ULA) of microphones. That is, the microphones are positioned on an axis with a uniform spacing between the microphones. The general expression of the microphone positioning in a ULA is given by

$$x_m = m \cdot d, \quad m = 0, 1, \ldots, M - 1$$

(3.1)

where $m$ denotes the microphone index; $x_m$ denotes the position of the microphone having index $m$, where $x_m = 0$ corresponds to the left-hand side of the array, i.e., the location where the microphone indexed by $m = 0$ is positioned; $d$ denotes the spacing between microphones, and $M$ is the array size, i.e., the number of microphones in the array, as demonstrated in Fig. 3.1(a) for the case of an even number of microphones and (b) for the case of an odd number of microphones. The center of the microphone array, i.e., the symmetry point, is set to be at $x = \left(\frac{M-1}{2}\right)d$.

The array microphones sample the incoming signals, collecting spatial samples that are then processed to attenuate signals from undesired directions and extract the signal from a desired direction. A spatial response of the microphone array is obtained via a beam (main-lobe) directed to the desired signal while nulls are directed to the undesired signals. As in Section 2.3, Figure 2.3 illustrates a beamformer design based on a linear array shown in Fig. 3.1.

Let $f(t, x)$ denote the set of signals sampled by the microphone array at time $t$, ex-
pressed by
\[ f(t, x) = [f(t, x_0), \ldots, f(t, x_{M-1})]^T, \] (3.2)

where \( x_m \) denotes the microphone position, \( t \) denotes the continuous-time variable, and \( T \) denotes the transpose operation. The beamformer output \( y(t) \) is expressed by:
\[ y(t) = \sum_{m=0}^{M-1} f(t, x_m) w_m^*, \] (3.3)

where \( w_m \) is a complex weight of the microphone having index \( m \), as shown in Fig. 2.3, and \( * \) denotes complex conjugation. The simplest beamformer design architecture has uniform weights, \( w_m = \frac{1}{M} \), \( m = 0, \ldots, M-1 \).

Let \( f_\omega(t, x) = e^{j\omega t} \) denote a plane wave propagating at angular frequency \( \omega \), and let \( \theta, \theta \in [-\pi/2, \pi/2] \), denote the direction of arrival (DOA) angle measured with respect to the broadside of the linear array, as shown in Fig. 2.3. The wave signals spatially sampled by the microphone array inputs are given by:
\[ f_\omega(t, x) = [f_\omega(t - \tau_0), \ldots, f_\omega(t - \tau_{M-1})]^T, \quad \tau_m = \frac{\sin(\theta) \cdot x_m}{c} \] (3.4)

where \( \tau_m \) is the propagation delay for the incoming signal and \( c = 340 \frac{m}{s} \) denotes the velocity of sound propagation in air. A value of \( \tau_m = 0, \forall m \) implies a DOA of \( \theta = 0, \)
i.e., a plane wave parallel to the array, propagating perpendicularly to the array. Let \( \kappa = \frac{\omega}{c} \sin(\theta) = \frac{2\pi}{\lambda} \sin(\theta) \) denote the wavenumber for plane waves in a locally homogeneous medium, where \( \lambda \) denotes the wavelength corresponding to the angular frequency \( \omega \), and let \( \mathbf{v}(\kappa) \) denote the array manifold vector \([37]\), featuring all of the spatial characteristics of the microphone array. Based on (3.4), and the definition of \( \kappa \) above, the manifold vector can be expressed as:

\[
\mathbf{v}(\kappa) = [e^{-j\kappa x_0}, \ldots, e^{-j\kappa x_{M-1}}]^T
\]  

Let \( P(\omega, \theta) \) denote the frequency and DOA dependent beamformer response. It is given by:

\[
P(\omega, \theta) = \sum_{m=0}^{M-1} e^{-j\omega \tau_m} w_m^* = \mathbf{w}^H \mathbf{v}(\omega, \theta),
\]

where \( \mathbf{w} = [w_0, \ldots, w_{M-1}]^T \) and \( H \) denotes Hermitian transpose operation.

Beampatterns are the main tool in assessing an array performance as they express the beamformer response for various frequencies and signal angles. Let \( D(\omega, \theta) \) denote the beampattern of a given beamformer. It is expressed by:

\[
D(\omega, \theta) = 20 \log_{10} |P(\omega, \theta)|,
\]

where \(|·|\) denotes the absolute value. For a given angular frequency \( \omega \), the beampattern \( D(\omega, \theta) \) is a function of the angle \( \theta \) and the beamwidth is measured in terms of \( \theta \). In this work, the beamwidth is measured between the two lowest values at both sides of the main lobe. So, when attainable, it is the beamwidth null to null, i.e., the width between zeroes of the main lobe \([37]\), given by \( 2 \sin^{-1}\left( \frac{2\pi c}{M \omega} \right) \).

The relation between the temporal frequency, \( f = \frac{\omega}{2\pi} \), the number of microphones and the beamwidth, \( \theta_{BW} \) can be obtained from the following expression:

\[
\sin\left( \frac{\theta_{BW}}{2} \right) = \frac{2\pi c}{M d \omega} = \frac{c}{M d f}
\]

Thus, as the temporal frequency \( f \) increases, the beamwidth decreases.

### 3.1.2 Temporal Filtering by an FIR Filter Array

Simple beamformers, as shown in Fig. 2.3, utilize a single weight for each microphone. This makes them ineffective when dealing with wideband signals, the signals of interest...
in speech and audio processing, as each of the signals is composed of various frequency components. In order to design a beamformer for wideband signals, the weight values in (3.6) must be altered for different frequencies in order to obtain the desired beamformer output. That is, the weights should be frequency dependent, i.e., in the form of \( w(\omega) = [w_0(\omega), \ldots, w_{M-1}(\omega)]^T \). This can be achieved by introducing discrete FIR filters [19,24].

FIR filters perform temporal filtering in order to compensate for the phase differences of the input wideband signals’ various frequency components. Fig. 2.4 shows the realization of frequency dependent weights by FIR filters connected to the microphone array.

Let \( y(t) \) denote the output of the FIR-based beamformer. It is expressed by:

\[
y(t) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(t - nT_s, x_m) \cdot w^*_{m,n}, \tag{3.9}
\]

where \( N \) denotes the number of FIR filter coefficients connected to each of the \( M \) microphones, \( w^*_{m,n} \) is the \( n \)-th coefficient of the FIR filter connected to the microphone indexed by \( m \), and \( T_s \) denotes the delay between adjacent filter elements. Given a complex plane wave signal, the beamformer output is given by:

\[
y(t) = e^{j\omega t} P(\omega, \theta) = e^{j\omega t} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} e^{-j\omega(\tau_m + nT_s)} \cdot w^*_{m,n}, \tag{3.10}
\]

Let \( v_s(\omega, \theta) \) denote a stacked array manifold vector of dimension \( M \cdot N \), where each subvector of dimension \( M \) represents the array manifold vector associated with a specific FIR filter coefficient in (3.10), i.e., the first subvector, \( [e^{-j\omega \tau_0}, \ldots, e^{-j\omega \tau_{M-1}}]^T \), is associated with the coefficient indexed by \( n = 0 \) and all of the array microphones, indexed by \( m = 0, \ldots, M - 1 \). Thus, it is denoted by \( v_0(\omega, \theta) \). The second subvector, \( [e^{-j\omega(\tau_0 + T_s)}, \ldots, e^{-j\omega(\tau_{M-1} + T_s)}]^T \), is associated with the coefficient indexed by \( n = 1 \), and all of the array microphones. Thus, it is denoted by \( v_1(\omega, \theta) \), and so on. This form of expression is called vector stacking, and \( v_s(\omega, \theta) \) is given by:

\[
v_s(\omega, \theta) = \begin{bmatrix} v_0(\omega, \theta) \\ v_1(\omega, \theta) \\ \vdots \\ v_{N-1}(\omega, \theta) \end{bmatrix}. \tag{3.11}
\]
CHAPTER 3. SPEECH ENHANCEMENT USING BEAMFORMERS

Given a ULA of \( M \) microphones, microphone spacing \( d \), and \( M \) FIR filters, each composed of \( N \) coefficients, and connected to a respective microphone. Then, from (3.10) and (3.4), the beamformer response can be expressed as:

\[
P(\omega, \theta) = \sum_{m=0}^{M-1} e^{-j\omega \tau_m} \sum_{n=0}^{N-1} e^{jn\omega T_s} \cdot w_{m,n}^* = w_s^H v_s(\omega, \theta),
\]

(3.12)

where \( w_s \) denotes the composite stacked weight vector of dimension \( M \cdot N \) created by vector stacking, having \( w_0 = [w_{0,0}, w_{1,0}, \ldots, w_{M-1,0}]^T \) as its first subvector, \( w_1 = [w_{0,1}, w_{1,1}, \ldots, w_{M-1,1}]^T \) as its second subvector, and so on. The design specification of the FIR filters will be addressed in Section 3.3. Figure 2.4. illustrates an FIR beamformer architecture.

The second summation in (3.12) can be expressed in terms of the Discrete-time Fourier transform (DTFT) of the \( m \)-th channel filter coefficients as follows:

\[
\sum_{n=0}^{N-1} e^{jn\omega T_s} w_{m,n}^* = W_m^* (e^{j\omega T_s}),
\]

(3.13)

where \( W_m(e^{j\omega T_s}) \) denotes the DTFT of \( w_{m,n} \) with the summation variable \( n \). Hence, the expression for \( P(\omega, \theta) \) in (3.12) can be written as:

\[
P(\omega, \theta) = \sum_{m=0}^{M-1} e^{-j\omega \tau_m} \cdot W_m^* (e^{j\omega T_s}).
\]

(3.14)

Since the channel filters have each a finite number of coefficients \( (N) \), their frequency responses can be represented by the Discrete Fourier transform (DFT) of size \( N \) of each filter coefficients. This corresponds to sampling \( \omega \) at \( \frac{2\pi}{NT_s} k, \ k = 0, \ldots, N - 1 \). Examining the beamformer response for frequencies corresponding to the bin centers defined by the DFT, we obtain:

\[
P(k, \theta) = \sum_{m=0}^{M-1} e^{-j\frac{2\pi \tau_m}{NT_s} k} \cdot W_m^* (e^{j\frac{2\pi}{NT_s} k}), \ k = 0, \ldots, N - 1,
\]

(3.15)

where \( k \) denotes the bin index of the DFT. This expression will be the basis for controlling the beamformer frequency response by specifying the frequency response of the FIR channel filters.

Many methods have been proposed for determining the FIR filter coefficients. The main goal of these methods is obtaining a desirable beamformer output while dealing
CHAPTER 3. SPEECH ENHANCEMENT USING BEAMFORMERS

with trade-offs between conflicting array performance measures. Traditionally, the performance measures of interest are the main lobe response, side-lobe levels, sensitivity to parameters mismatch, robustness to noise, directivity and aliasing. Frequency dilation based FIR beamformers [25] derive the FIR filter coefficients from a single reference frequency response, thus deriving all of the filter coefficients, which are microphone location dependent, from a single set of coefficients. This results in a constant beamwidth approximation in a wideband on one hand, but also in high side-lobe levels and low robustness on the other hand. Optimization-based beamformer designs such as the MVDR and LCMV beamformers [18], [20] force specific constraints on the beamformer weights and output. Therefore, these methods are chosen when a specific beamformer performance measure is needed to be optimized. The following section elaborates on the proposed method for determining the FIR filter coefficients that result in an approximate constant beamwidth beamformer for a wide range of input frequencies. Although the proposed method yields inferior beamwidth constancy in the lower frequency range than some methods, it offers improved array performance in terms of side-lobe attenuation, sensitivity and robustness, as well as a simple design method which is closed form.

3.2 Proposed FIR Design Algorithm

In this section, we present the proposed FIR design method that aims to obtain a constant beamwidth beamformer while maintaining lower levels of side-lobes compared to traditional methods [25]. The main feature of the proposed method is the determination of the FIR filter coefficients that provide a good approximation to a desired beamformer response. We consider real FIR coefficients since the signals of interest, i.e., acoustic signals, are real. For this purpose, we need first to determine the minimal frequency, i.e., the lowest frequency for which the desired beampattern beamwidth is feasible, given the array configuration. Then, the frequency response of each of the FIR filters is specified. To maintain a fixed delay by all the filters, they are designed to have a linear phase and use each the same number of coefficients. Finally, a post-summing FIR normalization filter is introduced at the output of the array to maintain a uniform gain of the beam at all the frequencies for which it is designed to have a constant beamwidth.
3.2.1 Frequency Range Determination for Constant Beamwidth Realization

The beamwidth of a given array is generally frequency dependent and gets smaller as the frequency increases. Thus, for a given desired beamwidth $\theta_{BW}$, and a given array of size $M$, there exists a minimal frequency, denoted $f_0$, for which $\theta_{BW}$ is realized. It is given from (3.8) by:

$$f_0 = \frac{c}{Md \sin \left( \frac{\theta_{BW}}{2} \right)},$$

(3.16)

where the desired beamwidth, i.e., the main lobe, extends from $-\frac{\theta_{BW}}{2}$ to $\frac{\theta_{BW}}{2}$. Hence, for input frequencies below $f_0$ the given array will produce a wider beamwidth than desired. Moreover, at higher frequencies than $f_0$, a lower beamwidth will be realized if no FIR filters are added.

To keep a constant beamwidth of $\theta_{BW}$, as the frequency of the impinging wave increases, we need to keep the product $M \cdot f$ constant. Since only a single frequency satisfies a constant $M \cdot f$ for a given $M$, $M$ must be changed to satisfy this relation for various $f$ values. We designate the microphones that satisfy a constant $M \cdot f$ as effective microphones. That is, as the wave frequency increases, the effective number of participating microphones in the array, denoted by $M_p$, is reduced by means of signal attenuation by the appropriate FIR filters. Since $M_p$ must be an integer, the product constancy will be satisfied exactly only for a set of frequencies, while between those frequencies it is approximately satisfied.

For a given array size, $M$, $M_p$ is limited to the range: $M_{\text{min}} \leq M_p \leq M$, $M_p = M - 2p, p = 1, 2, ..., \frac{M-M_{\text{min}}}{2}$. Theoretically, $M_{\text{min}} = 1$, but this value is not feasible since a single microphone is necessarily omni-directional and therefore cannot produce the desired beamwidth, assuming $\theta_{BW} < 2\pi$. Thus, the minimal value of $M_p$ is $M_{\text{min}} = 3$, for an odd number of microphones ($M$ odd), and $M_{\text{min}} = 2$ for an even number of microphones ($M$ even), maintaining symmetry around $x = \left( \frac{M-1}{2} \right) d$. Substituting the value of $M_{\text{min}}$ in (3.16) we obtain the maximal frequency, denoted $f_{\text{max}}$, for which we can still maintain the fixed beamwidth. For frequencies beyond $f_{\text{max}}$ the beam will be narrower than desired. Thus, the proposed approach can provide an approximate constant beamwidth in the range $[f_0, f_{\text{max}}]$. 
Different applications may desire a narrower range of frequencies for which the beamwidth is constant, say \([f_L, f_H]\), \(f_0 \leq f_L < f_H \leq f_{\text{max}}\), allowing a reduced complexity array, as will be discussed in Sections 3.3.2 and 3.3.4.

We continue here with the maximal range, i.e., \(f_L = f_0\) and \(f_H = f_{\text{max}}\). As the frequency increases within that range, the number of effective microphones needs to be reduced. This can be done only in decrement steps of 2, i.e., \(M_1 = M - 2\), \(M_2 = M - 4\), and so on, maintaining symmetry by effectively removing at each step one microphone from each end of the array and calculating a corresponding minimal frequency \(f_p\), \(f_L < f_p \leq f_H\), by replacing \(M\) by \(M_p\) in (3.16). Note that because \(M_p\) varies in decrements of 2 and not continuously, the beamwidth will fluctuate for frequencies between adjacent \(f_p\) values. This issue will be demonstrated and alleviated in the sequel.

### 3.2.2 FIR Filter Design

Consider a symmetric ULA and let \(P(f, \theta)\), denote its beamforming response. The proposed algorithm plugs in an FIR filter at each microphone input line, as in Fig. 2.4. Each FIR filter is of length \(N\) and can be represented by its frequency response, calculated by applying a length \(N\) DFT to the filters coefficients. This results in \(N\) frequency coefficients, one per each of the \(N\) frequency bins in \([0, \frac{1}{T_s})\), where \(T_s\) is the sampling interval of the input signal, i.e., the temporal sampling period of the system. The rationale behind the selection of \(T_s\) is discussed in this section. The design algorithm specifies custom frequency response coefficients of each filter needed to obtain a constant beamwidth beamformer.

**Basic Design**

The proposed approach for maintaining an approximate fixed beamwidth in the specified frequency range is to reduce the number of effective microphones by attenuating the signal from specific microphones as the frequency increases. This is achieved by letting the real FIR filters to be lowpass filters with cutoff frequencies set according to the values of \(f_p, p = 1, 2, \ldots, \frac{M-M_{\text{min}}}{2}\), as shown in Fig. 3.2, for \(M\) odd.

The expression in (3.15) gives the beamformer frequency response at \(N\) discrete fre-
quencies in terms of the DFT of the FIR filters coefficients. Let $W$ denote an $N \times M$ matrix denoted the filter array matrix based on this expression. The matrix is composed of $m$ columns, $m = 0, 1, 2, ..., M - 1$, where each column elements are given by $W^*_m \left( e^{j \frac{2\pi}{N} k} \right)$, corresponding to the DFT of the $m$-th channel filter coefficients. For the sake of simplified notation, and because the magnitude response of a causal linear-phase filter is the same as that of a zero-phase filter, we continue here with a zero phase filter. Since the magnitude responses of the considered filters have mirror-image symmetry with respect to half the sampling frequency, we partition $W$ into two blocks:

$$W = \begin{bmatrix} W_L \\ W_H \end{bmatrix}, \quad (3.17)$$

where both $W_L$ and $W_H$ are of size $L \times M$, and $L$ is the number of frequency bins in the lower band, i.e., up to half the sampling frequency, such that the columns of $W_L$ represent the magnitude response of the filters in the lower band, i.e., up to half the sampling frequency, and the columns of $W_H$ represent those of the upper band, up to the sampling frequency. Hence, we refer only to $W_L$ in the following design process.
Figure 3.2: The desired frequency responses of the FIR filters for the lower half of the frequency range, indexed by microphone index \( m = 0, \ldots, \frac{M-1}{2}, M \text{ odd} \). Due to the microphone array symmetry, the magnitude response specification for the filters connected to the second half of the array, i.e., for \( m = \frac{M+1}{2}, \ldots, M-1 \), is a mirror image of first half of the array. Each dot in the specified response is the value of the filter magnitude at a corresponding frequency bin. Since \( M \) is odd and due to symmetry, the two bottom rows of Fig. 3.2 represent all-pass FIR filters since \( M_{\text{min}} = 3 \). This means that three microphones are effective at all frequencies. For \( M \) even, only two microphones would be effective at all frequencies since \( M_{\text{min}} = 2 \) and therefore only the FIR filter connected to the microphone indexed by \( m = \frac{M-1}{2} \), i.e., the last row, would be an all-pass filter.
For an odd number of microphones ($M$ odd), $W_L$ is specified as:

$$W_L = \begin{pmatrix}
1 & 1 & 1 & 1 & \ldots & 1 & 1 & 1 & \ldots & 1 & 1 & 1 & 1 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
1 & 1 & 1 & 1 & \ldots & 1 & 1 & 1 & \ldots & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & \ldots & 1 & 1 & 1 & \ldots & 1 & 1 & 1 & 1 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
1 & 1 & 1 & 1 & \ldots & 1 & 1 & 1 & \ldots & 1 & 1 & 1 & 1 \\
0 & 1 & 1 & 1 & \ldots & 1 & 1 & 1 & \ldots & 1 & 1 & 1 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
0 & 1 & 1 & 1 & \ldots & 1 & 1 & 1 & \ldots & 1 & 1 & 1 & 0 \\
0 & 0 & 1 & 1 & \ldots & 1 & 1 & 1 & \ldots & 1 & 1 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
0 & \ldots & \ldots & \ldots & 0 & 1 & 1 & 1 & 0 & \ldots & \ldots & \ldots & 0 \\
\end{pmatrix}$$

Equation (3.18)

The columns of $W_L$ are the magnitude specifications of the FIR channel filters, as shown in Fig. 3.2. The row indices of $W_L$ correspond to the bin frequencies in the lower band. That is, the first row of $W_L$ in (3.18), corresponds to the frequency bin centered at $f = 0$, and the last row corresponds to half the sampling frequency. Since $M$ is odd in (3.18), the 3 central columns of $W_L$, corresponding to $M_{\text{min}} = 3$, are specified as all 1’s, i.e., allpass filters. For an even number of microphones $M$, the matrix in (3.18) would contain 2 central columns with all 1’s, corresponding to $M_{\text{min}} = 2$.

The first column of $W_L$ is the magnitude response of the FIR filter indexed by $m = 0$. The row corresponding to frequency $f_1$, indicates the cutoff frequency of that filter. For frequencies from $f_1$ up to $f_2$, $M_1 = M - 2$ microphones are effective. For the range $[0, f_1)$ all the $M$ microphones are effective. The second column of $W_L$ is the magnitude response of the FIR filter indexed by $m = 1$. Similarly, the frequency $f_2$ indicates the cutoff frequency of the filter associated with the second column. For frequencies from $f_2$ up to $f_3$ (not shown in (3.18)), $M_2 = M - 4$ microphones are effective, and so on. The number of effective microphones may reach $M_{\text{min}} = 3$ at a lower frequency than half the
sampling frequency, unless \( f_{\text{max}} = \frac{1}{2T_s} \). If \( f_H < f_{\text{max}} \), the number of effective microphones need not reach \( M_{\text{min}} \) but would be higher (matching \( f_H \)), so more columns in \( \mathbf{W}_L \) should be set all 1’s, i.e., all-pass filters. A similar situation will occur if the sampling frequency is set so that \( \frac{1}{2T_s} < f_{\text{max}} \), since \( f_H \) must then be set to half the sampling frequency, i.e., \( \frac{1}{2T_s} \).

In reality, a causal filter is eventually implemented by shifting the filter coefficients, resulting in a linear-phase causal filter with the same magnitude response. The basic design above improves the beamformer response, compared to the conventional one, in terms of beamwidth constancy, as demonstrated in Fig. 3.3(a) and Fig. 3.3(b). Yet, because \( M_p \) varies in steps of 2, the beamwidth does vary at frequencies that are between adjacent cutoff frequencies, causing fluctuations in the beamwidth, as clearly seen in Fig. 3.3(b). To reduce these fluctuations in beamwidth a modified design is proposed next.

Note also that for frequencies in between adjacent bins, the frequency response of the FIR filters also diverts from the desired ideal low-pass response because of the finite number of filter coefficients (\( N \)). By increasing \( N \), the latter effect is reduced but at the expense of increased computational complexity.

**Modified Design**

The beamwidth fluctuations noted above can be alleviated by modifying the specifications of some of the FIR filters in the array to include intermediate magnitude values, denoted as *smoothing magnitude coefficients*, in the transition band between the cutoff frequency of those filters to the cutoff frequency of the next filter in the array, as shown in Fig. 3.4. The corresponding form of the modified FIR magnitude response matrix \( \mathbf{W}_L \) is shown in (3.19). As in the basic design, we initially design a zero-phase FIR filter and eventually implement a causal filter by shifting the filter coefficients, resulting in a linear-phase causal filter with the same magnitude response. Let \( k \) denote the index corresponding to the \( k \)-th bin frequency. A row of \( \mathbf{W}_L \) consists of the responses of all the array filters at a given bin frequency, e.g., the row of \( \mathbf{W}_L \) corresponding to the frequency of \( f_0 \), is given the frequency bin index \( k_0 \). Due to the microphone array symmetry attributes, we define pairs of smoothing magnitude coefficients where every pair of coefficients is placed...
Figure 3.3: Array beampattern of (a) a conventional array without FIR filters. (b) an FIR array with magnitude response $W_L$, as given in (3.18). And (c) the modified design presented in the next section. The frequency range of $[0, 1/2T_s] = [0, 8]$ kHz is represented by the Y axis and is quantized into $N = 32$ frequency bins. The DOA range of $\theta = [-90, 90]$° is represented by the X axis. For display purposes, the frequency response at a higher resolution is computed from the FIR filter coefficients. The desired beampattern beamwidth is $\theta_{BW} = 30^\circ$ and the minimal frequency is $f_0 = 3412$ Hz.
at symmetrical microphone indices, i.e., at \( m = 0, M - 1 \), for frequency bins indexed by \( k_0 > k \geq k_2 \); \( m = 1, M - 2 \), for frequency bins indexed by \( k_2 > k \geq k_3 \), and so on. Although \( M \) effective microphones should be used in the transition band \( k_0 > k \geq k_1 \) and \( M - 2 \) effective microphones should be used in the transition band \( k_1 > k \geq k_2 \), we use \( M - 2 \) effective microphones for all bins in \( k_0 > k \geq k_2 \) to achieve the desired beamwidth. Utilizing \( M \) effective microphones in \( k_0 > k \geq k_1 \) would prevent controlling the beamwidth since there is no degree of freedom to widen the narrowing beamwidth in that range. We show below that these coefficients help in reducing the beamwidth fluctuations by adding a degree of freedom in controlling the beamwidth at those frequency bins that are between adjacent cutoff frequencies. Let \( W_{k,m} \) denote the smoothing magnitude coefficients that are shown in Fig. 3.4.

As depicted in Fig. 3.4, \( W_L \) takes the form shown in (3.19).

\[
W_L = \begin{pmatrix}
1 & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 & 1 \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
1 & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 & 1 \\
W_{k_0+1,0} & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 & W_{k_0+1,M-1} \\
W_{k_0+2,0} & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 & W_{k_0+2,M-1} \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
W_{k_2-1,0} & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 & W_{k_2-1,M-1} \\
0 & W_{k_2,1} & \cdots & 1 & 1 & 1 & \cdots & 1 & W_{k_2,M-2} & 0 \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & \cdots & 0 & W_{\left\lfloor \frac{N}{2} \right\rfloor, \frac{M-m}{2}-1} & 1 & 1 & 1 & W_{\left\lfloor \frac{N}{2} \right\rfloor, \frac{M-m}{2}-1} & 0 & \cdots & 0 \\
\end{pmatrix}
\]

Because the array is symmetric, the pair of smoothing coefficients in the relevant rows, i.e., rows corresponding to \( k > k_0 \), are set to have equal values. Furthermore, to allow a closed form solution for these coefficients, at most one pair of such coefficients appear in each row. The smoothing coefficients values are then derived as follows:

Let \( f_k \) denote the \( k \)-th frequency bin, where \( k = 0, \ldots, L-1 \). The beamformer response
Figure 3.4: Modified FIR magnitude responses for $M$ odd, where $W_{k,m}$ denotes the FIR filter magnitude response at the frequency bin indexed by $k$ and the microphone indexed by $m$.

at $f_k$, with the modified FIR filter array, is given, from (3.15) by:

$$
P(f_k, \theta) = \frac{1}{M_p + 2W_k} \cdot \left( M - \frac{M_p}{2} \right)_{-1} \sum_{m=\left\lfloor \frac{M-M_p}{2} \right \rfloor}^{\left\lfloor \frac{M-M_p}{2} \right \rfloor - 1} (e^{-j \frac{2\pi f_k}{c}} m d \sin(\theta)) +$$

$$+ W_{k, \left\lfloor \frac{M-M_p}{2} \right \rfloor - 1} \cdot e^{j \frac{2\pi f_k}{c}} \left( \left\lfloor \frac{M-M_p}{2} \right \rfloor - 1 \right) d \sin(\theta) + W_{k, M - \left\lfloor \frac{M-M_p}{2} \right \rfloor} \cdot e^{-j \frac{2\pi f_k}{c}} \left( M - \left\lfloor \frac{M-M_p}{2} \right \rfloor \right) d \sin(\theta),$$

(3.20)
where $W_k, \left\lfloor \frac{M-M_p}{2} \right\rfloor - 1$ and $W_{k, M - \left\lfloor \frac{M-M_p}{2} \right\rfloor}$ denote the real smoothing magnitude coefficients at the $k$-th frequency bin of the filter, $k_p$ denotes the $k$-th frequency bin corresponding to $f_p$, and $\frac{\sin(\theta) - m d}{c}$ is substituted for $\tau_m$ in (3.15).

As stated above, the two unknown coefficients in (3.20) are set to have the same value, so that the unknown coefficient value can be found by forcing the beampattern to have a null at $\theta_{\text{null}} = \pm \frac{\theta_{\text{BW}}}{2}$. Thus, we can find the smoothing magnitude coefficient $W_k, \left\lfloor \frac{M-M_p}{2} \right\rfloor - 1$ by setting $P(f_k, \theta_{\text{null}}) = 0$.

Since $\frac{1}{M_p + 2W_k, \left\lfloor \frac{M-M_p}{2} \right\rfloor - 1} \neq 0$, we get:

$$M - \left\lfloor \frac{M-M_p}{2} \right\rfloor - 1 \sum_{m = \left\lfloor \frac{M-M_p}{2} \right\rfloor} \frac{2\pi f_k}{c} m d \sin(\theta_{\text{null}}) =$$

$$= -W_k, \left\lfloor \frac{M-M_p}{2} \right\rfloor - 1 \left( e^{-j \frac{2\pi f_k}{c} \left( \left\lfloor \frac{M-M_p}{2} \right\rfloor - 1 \right) d \sin(\theta_{\text{null}})} + e^{j \frac{2\pi f_k}{c} \left( \left\lfloor \frac{M-M_p}{2} \right\rfloor - 1 \right) d \sin(\theta_{\text{null}})} \right)$$

Using Euler and other trigonometric identities, we get:

$$\sum_{m = 0}^{\left\lfloor \frac{M-p}{2} \right\rfloor} 2 \cos \left( \frac{2\pi f_k}{c} m d \sin(\theta_{\text{null}}) \right) =$$

$$= -W_k, \left\lfloor \frac{M-M_p}{2} \right\rfloor - 1 \cdot 2 \cos \left( \frac{2\pi f_k}{c} \left( \left\lfloor \frac{M-M_p}{2} \right\rfloor - 1 \right) d \sin(\theta_{\text{null}}) \right)$$

The final expression of the smoothing magnitude coefficient $W_k, \left\lfloor \frac{M-M_p}{2} \right\rfloor - 1$ is finally given by:

$$W_k, \left\lfloor \frac{M-M_p}{2} \right\rfloor - 1 = -\frac{\sum_{m = 0}^{\left\lfloor \frac{M-p}{2} \right\rfloor} \cos \left( \frac{2\pi f_k}{c} m d \sin(\theta_{\text{null}}) \right)}{\cos \left( \frac{2\pi f_k}{c} \left( \left\lfloor \frac{M-M_p}{2} \right\rfloor - 1 \right) d \sin(\theta_{\text{null}}) \right)}$$

(3.21)

Following the construction of $W_L$ in (3.19) and the placement of the new magnitude response coefficients calculated by (3.21), the discrete-time filter coefficients of each FIR filter are calculated by applying the IDFT to the corresponding column of $W$. The construction of $W$ requires the construction of $W_H$ from $W_L$ based on the mirror image symmetry of the magnitude response of $W_L$. Fig. 3.3(c) demonstrates the effect of the modification introduced in this section. It is clearly seen that the proposed modification
results in a better approximation of a constant beamwidth beamformer, compared to the results shown in Fig. 3.3(a) and Fig. 3.3(b).

### 3.2.3 Post Summing FIR Filter Design

Following the input FIR filter array, an additional normalization FIR filter is added. It is needed to achieve a uniform gain at all the frequencies in the range \([f_L, f_H]\), as the frequency responses of the different FIR filters are not all the same. It is inserted at the beamformer output.

Let \(\tilde{w}_k, k = 0, \ldots, N - 1\), denote the samples of the normalization filter frequency response at \(N\) discrete frequencies, i.e., the DFT coefficients of the post summing normalization FIR filter. The \(k\)-th DFT coefficient is set to:

\[
\tilde{w}_k = \frac{1}{M-1} \sum_{m=0}^{M-1} |W_{k,m}|
\]  

(3.22)

where \(W_{k,m}\) denotes the coefficient of \(W_L\) in (3.19) at the \(k\)-th row (frequency bin) and at the column indexed by \(m\) (the microphone index). For every frequency bin, the normalization filter sums the absolute values of the frequency response coefficients for the microphone filters from \(m = 0\) to \(m = M - 1\). To achieve a uniform gain also at frequencies that are in between the \(N\) bin frequencies, more than \(N\) coefficients for the normalization filter are needed. The additional coefficients values can be computed by first obtaining a higher resolution frequency response of the FIR filters by zeros padding, as mentioned earlier, and then applying normalization as in (3.22) for the additional frequencies. The normalization filter coefficients are then obtained by applying the IDFT to the frequency response coefficients. Fig. 3.5 illustrates the proposed complete FIR filter based beamformer structure.

### 3.2.4 Comments

If it is desired that the microphone array should not pass signal frequencies below \(f_0\), because the beamwidth at those frequencies is wider than desired, one can use FIR filters of the bandpass type, having their lower cutoff frequency at \(f_0\). However, bandpass filters are generally more complex to implement, therefore, the lowpass type would be
usually preferred. Note that the bandpass filters could be designed to attenuate also signal frequencies above $f_{\text{max}}$, if it is below $\frac{1}{2T_s}$.

In the design presented above we assumed that $[f_L, f_H] = [f_0, f_{\text{max}}]$. However, if $f_L > f_0$, one can use less microphones in the array so as to match $f_0$ to the desired $f_L$.

The proposed algorithm’s main stages of array configuration and FIR filter design are summarized in Table 3.1.

### 3.3 Experimental Results

In this section, we demonstrate the performance and advantages of the proposed algorithm via several experiments. We compare the acquired results to three different beamforming algorithms: (i) Parra’s algorithm [20], where a least-squares optimal basis transform is calculated, decoupling the frequency response from the spatial response. (ii) The algorithm by Tourbabin et al. [63], where a real-valued solution to the maximum directivity optimization problem is presented. (iii) Doblinger’s algorithm [16], where second-order cone programming based optimization is used together with sensor calibration. All of the simulated microphone arrays are configured with $M = 11$ omnidirectional microphones. Moreover, each input channel is connected to a $N = 32$ coefficient long FIR filter and the sampling frequency of the system is $f_s = 16$ kHz, where $f_s = \frac{1}{T_s}$. In Section 3.4.1 we run simulations for a uniform linear array with $d = 3.5$ cm spacing. Section 3.4.2 provides simulation results when microphone gain imperfection are induced by a random signal amplitude difference of up to 15% between the microphones. In Section 3.4.3 we provide simulation results for microphone array spacings perturbations induced by a random
Array Configurations:

1. Calculate the minimal and maximal frequency values $f_0$ and $f_{\text{max}}$, given the array specifications (number of array microphones $M$, desired beamwidth $\theta_{\text{null}}$ and spacing $d$) using (3.16).

2. Calculate $f_p$, $f_L < f_p \leq f_H$, by replacing $M$ by $M_p$ in (3.16);
   
   $M_{\text{min}} \leq M_p \leq M - 2, \quad p = 1, 2, \ldots, (M - M_{\text{min}})/2$.

Input FIR Filters:

3. Construct the matrix $W_L$ as in (3.19).

4. Construct the matrix $W_H$ on the basis of the mirror image property of $W$.

5. Construct the matrix $W$ by concatenating $W_L$ and $W_H$.

6. Calculate the smoothing magnitude coefficients $W_k, \left\lfloor \frac{M - M_p}{2} \right\rfloor - 1$ using (3.21);
   
   $k = 0, \ldots, L - 1$.

Post Summing FIR Filter:

7. Calculate the post summing FIR filter coefficients using (3.22).

microphone positioning error of up to 10% from the nominal microphone positionings.
3.3.1 ULA Experiments

In this section, the spacing between each of the \( M = 11 \) microphones is \( d = 3.5 \) cm, the desired beamwidth is \( 30^\circ \), the frequency range is \([f_L, f_H] = \left[ f_0, \frac{1}{2T} \right] = [3400, 8000]\) Hz. The simulation results are shown in Fig. 3.6 where the null-to-null beamwidth is bounded by the two black arrows in Fig. 3.6(a), the side-lobes are marked by the two white arrows in Fig. 3.6(b) and the aliasing phenomena is marked by the two gray arrows in Fig. 3.6(c).

![Figure 3.6: Simulation results for ULA frequency response of (a) the proposed algorithm. (b) Parra’s algorithm. (c) Tourbabin’s algorithm. (d) Doblinger’s algorithm.](image)

In the following sections, we elaborate on the performance of the various methods.
in beampattern beamwidth constancy, side-lobe attenuation, aliasing effects attenuation, sensitivity, microphone gain imperfections and array spacings perturbations.

**Beampattern Beamwidth**

In Fig. 3.7 the effective beampattern beamwidths of the various methods are compared. The beamwidths are compared by their error relatively to the desired beamwidth $\theta_{BW} = 2\theta_{null}$, i.e., the difference between the null-to-null beamwidth of each algorithm and the desired beamwidth in the wideband.

![Figure 3.7: Beampattern beamwidth error of the proposed algorithm (blue square). Parra’s algorithm (red circle). Tourbabin’s algorithm (green solid). Doblinger’s algorithm (black triangle). conventional ULA (cyan star).](image)

The beamwidth constancy can be measured by the mean absolute error (MAE) between the beamwidth received by the various methods and the desired beamwidth in the examined frequency range. The received MAE values were $1.7^\circ$, $1.7^\circ$, $7.7^\circ$, $1.4^\circ$ and $10^\circ$ for the proposed, Parra’s, Tourbabin’s, Doblinger’s and ULA delay and sum methods, respectively. The reason for the somewhat inferior results of the proposed method, in comparison to two of the examined methods, is that these methods use optimization. However, the other methods do not offer direct control over the desired beamwidth as in
the proposed method and have a more complex design algorithm because of optimization. The proposed method also offers better side-lobe attenuation and aliasing attenuation, as well as more robust performance when microphone gain imperfections and array spacings perturbations are introduced, as shown in the next sections.

**Side-lobe Attenuation and Output Gain**

To compare the side-lobe attenuation performance, Fig. 3.8 presents the beampatterns obtained for the different methods. The curves shown in Fig. 3.8 represent the beampatterns at different frequencies in the frequency range of [3400, 8000] Hz with increments of 230 Hz. The proposed algorithm demonstrates better side-lobe attenuation without compromising main-lobe beamwidth and output gain. The side-lobe levels are marked by the red dashed lines. A side-lobe attenuation of 13.42 dB is obtained for the proposed algorithm in Fig. 3.8(a), while side-lobe attenuations of 7.71 dB, 10.78 dB, and 14.72 dB are obtained for the other algorithms, as seen in Fig. 3.8(b)-(d), respectively. The primary reason for the better side-lobe attenuation performance in the proposed algorithm is the explicit calculation of the beampattern beamwidth for every frequency bin, shown in (3.18)–(3.21), together with the attenuation caused by the decreasing number of effective microphones, causing most of the power to stay within the main-lobe and not leak to the side-lobes. The algorithm proposed by Doblinger [16] minimizes the side-lobe levels at the expense of the main-lobe level, which is lower than 0 dB some frequencies, as shown in Fig. 3.8(d). The other methods do not take side-lobe levels attenuation performance into consideration. In Tourbabin’s algorithm [63], lowering the side-lobe levels by using different cost functions decreases the performance of the algorithm’s main objective.

In Fig. 3.8(a) and Fig. 3.8(c) the beampatterns have a uniform gain due to normalization factors. The proposed algorithm utilizes an output FIR post-summing normalization filter and Tourbabin’s algorithm utilizes normalization constants for obtaining a uniform gain beampattern. Parra’s and Doblinger’s algorithms in Fig. 3.8(b) and Fig. 3.8(d) do not use normalizations factors and therefore yield non-uniform gain beampatterns. In Parra’s algorithm, the phenomena is more prevalent at low frequencies, while in Doblinger’s algorithm it is more noticeable at higher frequencies. Output FIR normalization filters have been utilized as part of the FIR filter design in several beamforming design meth-
Figure 3.8: Simulation results for ULA beampatterns at different frequencies of (a) the proposed algorithm. (b) Parra’s algorithm. (c) Tourbabin’s algorithm. (d) Doblinger’s algorithm. The side-lobe levels near the main-lobe are marked by the red dashed lines and the side-lobe levels at the edges, due to aliasing, are marked by the black dash-dot lines.

Aliasing Attenuation

Spatial aliasing occurs when the array microphones sample the impinging signals from different locations not densely enough, i.e., the microphone spacing $d$ is too large. This causes sources at different locations to have the same manifold vector (3.5). As a result, these sources locations cannot be uniquely determined based on the received array signals.
Avoiding aliasing requires to satisfy the condition \( d < \frac{\lambda}{2} \), where \( \lambda \) denotes the wavelength corresponding to the angular frequency \( \omega \). On the other hand, larger \( d \) values are often used for spatial resolution improvement. Although spatial resolution can be improved by increasing the number of array microphones, hardware costs and array size constraints are the leading consideration in most real world applications. Aliasing is a phenomenon whose effect is stronger at higher frequencies, inducing high side-lobe levels at \( \pm 90^\circ \). The proposed algorithm provides better attenuation of the high side-lobe levels at \( \pm 90^\circ \), caused by aliasing, than most of the other examined methods, as seen in Fig. 3.8. Parra’s algorithm offers slightly better aliasing attenuation because most of the energy outside of the main-lobe goes to side-lobes adjacent to the main-lobe.

**Sensitivity**

A major challenge in practical beamformer applications is the potential sensitivity to mismatches between the actual array attributes and the model used to derive the desired beamformer. In practical applications, mismatches can occur either by array spacings perturbations, production faults or filter perturbations. The sensitivity function often used as a criterion for assessing the affect of mismatches on the array response is defined in [37] by:

\[
T_{se} = A_w^{-1} = \|w\|^2,
\]

(3.23)

where \( A_w^{-1} \) is the inverse expression of the white noise gain given by \( A_w = \text{SNR}_{out}(k) / \text{SNR}_{in}(k) \) and \( w \) is the weight vector corresponding to all of the FIR filter channels in the \( k \)-th frequency bin. Therefore, as the white noise gain increases, the sensitivity decreases and the array would be more robust to mismatch. In Fig. 3.9 the sensitivity as a function of frequency of the various algorithms is compared. The proposed algorithm yields better results for a small part of the frequency spectrum, \((3400 - 4400) \text{ Hz}\), thanks to the decreasing number of effective microphones with frequency and the normalization FIR filter calculated in (23). At the higher part of the spectrum \((4400 - 8000) \text{ Hz}\), the proposed algorithm produces inferior results only to the method proposed by Tourbabin since the latter minimizes the sensitivity for higher frequencies [63].
CHAPTER 3. SPEECH ENHANCEMENT USING BEAMFORMERS

3.3.2 Micophone Gain Imperfections

For simulating microphone gain imperfections of a practical microphone array, we ran 10 simulations with a random microphone gain deviation of up to ±15% of the original gain and averaged the results. Figure 3.10 compares the effective beampattern beamwidths of the various algorithms before and after the gain imperfection introduction. This is done by subtracting the beamwidth of the gain mismatched array from the beamwidth of the original array shown in Fig. 3.9, for every method.

Table 3.2 shows the side-lobe attenuation and the MAE from the desired beamwidth of the various methods before and after the microphone array gain perturbation. The proposed method yields the best results for side-lobe attenuation and aliasing effects attenuation when array gain perturbations are introduced. Moreover, the array perturbations have the smallest effect on the beamwidth of the proposed method since the mean absolute error has the smallest deviation before and after the gain perturbations.
3.3.3 Array Spacings Perturbations

Similarly to the previous section, we simulate a practical microphone array considering array spacings perturbations due to inaccuracies in microphone positioning. We ran 10 simulations with random spacing deviations of up to ±10% in inter-microphones spacing. We computed the beampattern as a function of frequency using (3.15), and then averaged the results. Figure 3.11 compares the change in the effective beamwidths of the various algorithms, before and after introducing the array spacings perturbations. This is done
by subtracting the beamwidth of the spacing mismatched array from the beamwidth of the original array shown in Fig. 3.7, for every method. It is seen that the array spacings perturbations have very little effect on the effective beamwidth of the proposed method, compared to the other examined methods, as also seen by the MAE values in the left column of Table 3.3.

The scenarios of array spacings perturbations and microphone gain imperfection were examined in [64]. Although an optimal solution has been successfully obtained via a minimax criterion, the solution runtime is relatively long and suitable mainly for small arrays.
Table 3.3 also shows that the side-lobe levels of the various methods before and after the microphone array spacings perturbations. The proposed method is seen to yield the best results for side-lobe and aliasing attenuation, when array spacings perturbations are introduced.

Table 3.3: Side-lobe attenuation of the various methods before and after the microphone array spacings perturbation. The percent of degradation for every method is given in brackets.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>13.42/12.19 (9.17%)</td>
<td>15.36/13.1 (14.7%)</td>
<td>1.7/1.73 (1.76%)</td>
</tr>
<tr>
<td>Parra’s</td>
<td>7.71/5.85 (24.12%)</td>
<td>16.42/13.54 (17.54%)</td>
<td>1.7/1.81 (6.47%)</td>
</tr>
<tr>
<td>Tourbabin’s</td>
<td>12.92/9.84 (23.84%)</td>
<td>10.78/8.05 (25.32%)</td>
<td>7.7/8.94 (16.1%)</td>
</tr>
<tr>
<td>Doblinger’s</td>
<td>20.14/11.21 (44.34%)</td>
<td>14.72/6.21 (57.81%)</td>
<td>1.4/1.03 (26.43%)</td>
</tr>
</tbody>
</table>

3.3.4 Experimental Results Conclusions

Although the proposed method achieved inferior results to some methods in terms of the MAE in beamwidth, in some part of the frequency band, Fig. 3.8 shows that the proposed method yields better side-lobe and aliasing attenuation. The proposed method provides also a better performance, in terms of the sensitivity criterion, than the other examined algorithms in the lower part of the spectrum, as shown in Fig. 3.9. Finally, according to Fig. 3.10, Fig. 3.11, Table 3.2, and Table 3.3, when microphone gain and positioning imperfections were introduced, the proposed method provided the most robust results in beamwidth error, having the smallest change from the nominal performance.

3.4 Conclusions

A new method for the design of an FIR-based constant beamwidth beamformer has been proposed. First, an input FIR filter array is designed to obtain a fixed main-lobe width in a wide frequency range. Then, an output FIR normalizing filter is designed to keep a fixed gain over that band. The proposed design method achieves a good approximation to a constant beamwidth output response, with better side-lobe and aliasing attenuation, as well as robustness to array mismatches, as compared to the other examined methods. When inducing random microphone gain and array spacings perturbations, the proposed
algorithm produces steady beampatterns throughout the specified frequency band. The new method is based on designing the FIR filter array such that its frequency response reduces the number, of what we call, ”effective microphones”, when signal frequency increases, affecting an approximate fixed beamwidth in a wide range of frequencies. Simulations of the proposed beamformer also demonstrate low computational complexity due to its simple closed form FIR filter design procedure. A possible future research direction could be the application of the proposed design method to other array geometries besides ULAs.
Chapter 4

Voice Activity Detection in Presence of Transient Noise Using Spectral Clustering and Diffusion Kernels

In this chapter, we present a voice activity detection algorithm based on spectral clustering and diffusion maps and evaluate its performance. This chapter begins with the problem formulation in Section 4.1, where we also elaborate the discussion on the theory of the proposed method. Section 4.2 presents the performance and advantages of the simulation results of the proposed VAD method. Finally, Section 4.3 concludes the proposed VAD research and reviews possible future research directions.

4.1 Problem Formulation

In this section, we elaborate the discussion on the theory of the proposed algorithm. The proposed VAD system utilizes spectral clustering [32] and diffusion kernel [71] methods in order to find a novel way of calculating a similarity matrix.

Let $x_{sp}(n)$ denote a speech signal and let $x_{tr}(n)$ and $x_{st}(n)$ denote additive interfering transient and stationary noise signals, respectively. The microphone input signal is given by

$$ y(n) = x_{sp}(n) + x_{tr}(n) + x_{st}(n). $$

(4.1)
The proposed VAD algorithm generates a cloud of samples for each short frame (approximately 20 ms long) of the input signal, calculates MFCCs for each sample in order to from an MFCC matrix. Then, the algorithm calculates a covariance matrix for each frame. The covariance matrix adds additional factor of similarity between frames, which is utilized for the calculation of a similarity matrix.

4.1.1 Sample Cloud

Given an audio input signal, the algorithm divides it into $N$ frames, approximately 20 ms long. A moving window of size $M$ is then utilized in order to generate $i$ new samples of the frame. Every sample of the frame is regarded as an iteration of the generation process for the sample cloud given by

$$\hat{y}_j(n - k) = y(n) \cdot w_M(n - k - M \cdot (j - 1) + \frac{i}{2}), \ k = 1, ..., i \quad (4.2)$$

where $\hat{y}_j(n - k)$ is the generated sample cloud of the $j$-th frame (out of $N$), $y(n)$ is the original signal, $w_M$ is the moving window of size $M$, $i$ is the desired number of samples per frame and $k$ is the current iteration index. For every iteration $k$ of the sample cloud, the algorithm calculates $m$ MFCCs. MFCCs are coefficients that form a representation of the short-term power spectrum of a sound. MFCC is based on a linear cosine transformation of a log power spectrum on a non-linear Mel scale frequency, thus convenient for human auditory applications. An $m \times i$ matrix of MFCCs is created in this fashion for each frame of the $N$ frames. Finally, for each of the MFCC matrices the algorithm calculates a $m \times m$ covariance matrix.

Let $X_m^j$ be the $m \times i$ matrix of MFCCs of the $j$-th frame. The covariance matrix for the $j$-th frame is given by

$$\Sigma = \mathbb{E} \left( (X_m^j - \mathbb{E} (X_m^j)) (X_m^j - \mathbb{E} (X_m^j))^T \right) \quad (4.3)$$

Where $\mathbb{E}$ denotes the expected value of a matrix. With the covariance matrix $\Sigma$, we find the correlation between samples in the sample cloud.
4.1.2 Similarity Matrix

The most important part of the proposed VAD algorithm is the similarity matrix. The similarity matrix is utilized in order to effectively cluster the data and label it as speech or non-speech. Given an audio input signal composed of a combination of speech, stationary noise and transient noise components (i.e., $x_{sp}(n)$, $x_{st}(n)$ and $x_{tr}(n)$, respectively), we choose absolute value of MFCCs and the arithmetic mean of the log-likelihood ratios for the individual frequency bins as the feature space, as in [32].

Let $Y_m(t, k) \ (t = 1, ..., N; k = 1, ..., K_m)$ and $Y_s(t, k) \ (t = 1, ..., N; k = 1, ..., K_s)$ be the absolute value of the MFCC and the STFT coefficients in a given time frame, respectively. Both MFCC and STFT coefficients are computed in $K_m$ and $K_s$ frequency bins, respectively. Then, each frame is represented by a column vector of dimension $(K_m + 1)$ defined as follows

$$Y(:, t) = \begin{bmatrix} Y_m(:, t) \\ \Lambda_t \end{bmatrix} \quad (4.4)$$

where $Y_m(:, t)$ denotes the absolute value of MFCCs in a specific time frame $t$. $\Lambda_t$ denotes the arithmetic mean of the log-likelihood ratios for frame $t$. The expression combines various statistical calculations on the noise in the training stage as well as STFT coefficients of the input audio signal. $\Lambda_t$ is given by

$$\Lambda_t = \frac{1}{K_s} \sum_{k=1}^{K_s} \left( \frac{\gamma_k(t) \xi_k(t)}{1 + \xi_k(t)} - \log (1 + \xi_k(t)) \right) \quad (4.5)$$

where $\xi_k(t) = \lambda_s(t, k)/\lambda_N(t, k)$ and $\gamma_k(t) = |Y_s(t, k)|^2/\lambda_N(t, k)$ denote the a-priori and a-posteriori SNR [42], respectively. $\lambda_s(t, k)$ denotes the variance of speech signal in the $k$-th frequency bin of the $t$-th frame and $\lambda_N(t, k)$ denotes the variance of stationary noise in $t$-th time frame and $k$-th frequency bin.

Combining (4.3)-(4.5), we can now define the expression for the similarity matrix

$$W_\theta^t(i, j) = \exp \left( \sum_{p=-P}^{P} -\alpha_p Q(i + p, j + p) \right) \quad (4.6)$$
\[ Q(i, j) = \left[ Y_m^\ell(:, i) \left( 1 - \exp\left(-\Lambda_i^\ell/\epsilon\right) \right) - Y_m^\ell(:, j) \left( 1 - \exp\left(-\Lambda_j^\ell/\epsilon\right) \right) \right] \cdot \left( \Sigma_i^\ell + \Sigma_j^\ell \right)^\dagger \cdot \left[ Y_m^\ell(:, i) \left( 1 - \exp\left(-\Lambda_i^\ell/\epsilon\right) \right) - Y_m^\ell(:, j) \left( 1 - \exp\left(-\Lambda_j^\ell/\epsilon\right) \right) \right]^T \]

\[(4.7)\]

where \( \theta = [\epsilon, \alpha_{-P}, \alpha_{-P+1}, \cdots, \alpha_{P-1}, \alpha_P] \in \mathbb{R}^{2P+2} \) is a vector of system parameters, \( Y_m^\ell(:, i), \Lambda_i^\ell \) and \( \Sigma_i^\ell \) are the absolute value of the MFCC, the arithmetic mean of the log-likelihood ratio and the covariance matrix of the \( i \)-th frame in the \( \ell \)-th sequence, respectively, \( \epsilon \) is the kernel width obtained during the training stage and \( \dagger \) denotes the pseudo-inverse of a matrix. The main motivation behind the proposed similarity matrix calculation in (4.7) is finding a model for the signal generating system, i.e. the speech system of the speaker. With the new representation, we gain a smaller degree of freedom for the system model. We tag the system as a "black box" and try to find a model for the system by viewing its' outputs. In fact, a second order approximation is applied on the parameters in order to receive random Gaussian perturbations. A covariance matrix is then calculated and used in order to express a Jacobian matrix. Finally, the Jacobian is used in order to find a similarity matrix. In order to calculate the pseudo-inverse of the expression \( \Sigma_i^\ell + \Sigma_j^\ell \) in (4.7), we use the first three largest eigenvectors received in singular vector decomposition (SVD).

Let \( \Sigma \) be a covariance matrix as in (4.3), applying SVD yields

\[ \Sigma^\dagger = VS^\dagger \Delta^T \]

\[(4.8)\]

Where \( \Delta \) is an orthogonal matrix of size \( 3 \times N \), the columns of \( \Delta \) are the eigenvectors of \( \Sigma \Sigma^T \). \( S \) is a diagonal matrix at the same size of \( \Sigma \), its' values are the square roots of the non-zero eigenvalues of both \( \Sigma \Sigma^T \) and \( \Sigma^T \Sigma \). \( V \) is an orthogonal matrix, the same size of \( \Delta \). The columns of \( V \) are the eigenvectors of \( \Sigma^T \Sigma \).

### 4.1.3 Training Stage

The training algorithm in our research is based on [32]. Given databases of clean speech, transient noise and stationary noise signals, We choose \( L \) different signals from each database. Without loss of generality, we take the \( \ell \)-th speech signal, transient noise and
CHAPTER 4. SPEECH ENHANCEMENT USING VOICE ACTIVITY DETECTORS

stationary noise and combine them as in Fig. 4.1. We assume that all of these signal are of the same length of \( N \). With the new database and by utilizing (4.4) and (4.5), we extract the feature matrix \( Y^\ell_1, Y^\ell_2, Y^\ell_3 \). By concatenating the feature matrix, we build the \( \ell \)-th training sequence, \( Y^\ell \), as shown in Fig. 4.1. For each frame \( t \), in the training sequence \( \ell \) we compute an indicator matrix \( C^\ell_t \) in order to indicate a speech containing frame. For further discussion, see [32].

Next, we define a kernel which preserves the similarity between points, as the similarity matrix in (4.6) and (4.7). This metric guarantees small similarity between two frames.
of different classes, i.e., speech and transient noise, even if they are very similar to each other (in the Euclidean sense). This is enabled due the large distance between neighboring frames. Upon defining the parametric weight function, the parameters can be obtained by solving the following optimization problem [44]:

$$\theta^{\text{opt}} = \arg \min_{\theta} \frac{1}{L} \sum_{\ell=1}^{L} F(W_\ell^\theta, C_\ell) \quad (4.9)$$

$$F(W, C) = \frac{1}{2} \| \Upsilon \Upsilon^T - D^{1/2}C(C^T D C)^{-1} C^T D^{1/2} \|_F^2 \quad (4.10)$$

where $\Upsilon$ is an approximate orthonormal basis of the projections on the second principal subspace of $D^{-1/2} W D^{-1/2}$.

Let $W_{\theta^{\text{opt}}}$ be the similarity matrix of the $\ell$-th training sequence and $U_\ell$ be a matrix consisting of the two eigenvectors of $D^{\ell-1/2} W^\ell D^{\ell-1/2}$ corresponding to the first two largest eigenvalues, where $D$ is a diagonal matrix whose $i$-th diagonal element equals to $\sum_{j=1}^{N} W(i, j)$. We then define $U$ as the column concatenation of $U_1$ through $U_L$. $U$ is a new representation of the training data such that each row of $U$ corresponding to a specific training frame. For further information, see [32].

We use Gaussian mixture modeling to model each cluster, i.e., label as speech presence or absence, with a different Gaussian Mixture Model (GMM). This means that we model the low dimensional representation of the original data using two different GMMs, one for each cluster. Let $f(\cdot; \mathcal{H}_0)$ and $f(\cdot; \mathcal{H}_1)$ be the probability density function corresponding to speech absence and presence frames, respectively. The likelihood ratio for each labeled frame $t$ is then obtained by

$$\Gamma^{\text{train}}_t = \frac{f(U(t, :); \mathcal{H}_1)}{f(U(t, :); \mathcal{H}_0)} \quad (4.11)$$

where $U(t, :)$ is the $t$-th row of the matrix $U$, and $\mathcal{H}_1$ and $\mathcal{H}_0$ are the speech presence and absence hypotheses, respectively.

### 4.1.4 Testing Stage

The main goal of the testing stage is to cluster the unlabeled data and decide whether a given unlabeled frame contains speech or not. In order to compute the likelihood ratio for a new unlabeled frame, [32] utilizes GMM to model the eigenvectors of normalized
CHAPTER 4. SPEECH ENHANCEMENT USING VOICE ACTIVITY DETECTORS

Laplacian matrix.

\[ \Gamma_{\text{test}}^t = \frac{f(\tilde{U}(t,:); \mathcal{H}_1)}{f(U(t,:); \mathcal{H}_0)} \quad (4.12) \]

where \( \tilde{U}(t,:) \) is the \( t \)-th row of the new representation of the unlabeled data in terms of eigenvectors of the normalized Laplacian of the similarity matrix. In [40] it was shown that using the information supplied by neighboring frames can improve the performance of VAD algorithms. The improvement is enabled due to the fact that frames containing speech signal are usually followed by a frame that contains speech signal as well. In the contrary, transient signals usually last for a single time frame. Using this fact, the decision rule for an unlabeled time frame is obtained by:

\[ VA_t = \sum_{j=-J}^{J} \Gamma_{t+j} \mathcal{H}_1 \geq \mathcal{H}_0 \quad T_h \quad t = 1, 2, \cdots, T \quad (4.13) \]

where \( T_h \) is a threshold which controls the tradeoff between probability of detection and false alarm. Increasing (decreasing) this parameter leads to a decrease (increase) of both the probability of false alarm and the probability of detection. Both the training and testing stages are summarized in Table 4.1. The block schemes of both learning and testing stage are depicted in Fig. 4.1.

4.2 Simulations Results

In this section we demonstrate the performance and advantages of the proposed VAD algorithm via several simulations. We compare the results acquired to the results of the VAD algorithm proposed in [32]. We run the simulations for various SNR values, stationary noises and transient noises. We configure the number of sequences to 4 training sequences and 20 testing sequences. Speech signals are taken from the TIMIT database [72], and transient noise signals are taken from [73]. The sampling frequency is set to 16kHz. Furthermore, we pick a window of size 512 for STFT calculations, \( m = 14 \) mel-frequency bands, \( M = 257 \) as the size of the moving window to utilize in order to create the sample cloud and \( i = 45 \) as the number of samples for each frame. The graphs of probability of detection, \( P_d \), vs. probability of false alarm, \( P_{Fa} \), are depicted in Fig 4.2.
We use identical experiment conditions with both the proposed algorithm and [32] in every simulation. In Fig 4.3, we provide the clustering results, i.e., the U space representation where speech labeled data is marked with blue rings and non-speech data is marked with red crosses. The proposed VAD algorithm has superior performance in the entire SNR range, especially for low SNR values. Moreover, the proposed algorithm performs better in cases of very small training sets.

Figure 4.2: Probability of detection ($P_d$) versus probability of false alarm ($P_{fa}$), for various noise environments. (a) Babble noise with SNR of 5dB, and transient noise of door knocks. (b) White noise with SNR of 5dB, and transient noise of door knocks. (c) White noise with SNR of 10dB, and transient noise of typing. (d) Babble noise with SNR of 5dB, and transient noise of typing.
Figure 4.3: Clustering results of training stage for babble noise with SNR of 5dB, and transient noise of typing of (a) the proposed algorithm, and (b) the algorithm proposed in [44].

4.3 Conclusions

We have presented a VAD algorithm based on spectral clustering and diffusion kernel methods. The main challenge was providing good results in presence of environmental noise and transient noise in particular. The key features of the proposed algorithm are the covariance matrix calculations via sample clouds and the novel similarity matrix computations. We demonstrated better results compared to a work that has already been proven to be superior to conventional methods of dealing transient noises, especially in cases of low SNR and small data bases. The goal in the near future would be trying to improve the algorithm’s results using enhanced features. Another possible research direction would be choosing better parameters for the GMM.
CHAPTER 4. SPEECH ENHANCEMENT USING VOICE ACTIVITY DETECTORS

Table 4.1: Proposed Voice Activity Detection Algorithm Based on Spectral Clustering Method.

<table>
<thead>
<tr>
<th>Learning algorithm:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Construct a training data set consisting of</td>
</tr>
<tr>
<td>(L) training signals ({Y^{\ell} \in \mathbb{R}^{K_{m+1} \times 3N}; \ell = 1, \ldots, L})</td>
</tr>
<tr>
<td>and (L) indicator vectors ({C^{\ell} \in \mathbb{R}^{3N_{m+1}}; \ell = 1, \ldots, L}).</td>
</tr>
<tr>
<td>2. Solve the optimization problem given in (4.9), to find</td>
</tr>
<tr>
<td>the optimum value of the parameters (i.e. (\theta^{opt})).</td>
</tr>
<tr>
<td>3. Construct (U) by concatenation of (U^1) through (U^L)</td>
</tr>
<tr>
<td>(K) largest eigenvectors of (D^{-1/2}WD^{-1/2}).</td>
</tr>
<tr>
<td>4. Fit a GMM model to the rows of (U) for</td>
</tr>
<tr>
<td>each cluster (see (32)).</td>
</tr>
<tr>
<td>Output:</td>
</tr>
<tr>
<td>(U, f(\cdot; H_1)) and (f(\cdot; H_0))</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Testing Procedure:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Let (z_t(n)) be the test sequence and (Z_t)</td>
</tr>
<tr>
<td>the feature matrix of (t)-th unlabeled frame obtained by (4.5).</td>
</tr>
<tr>
<td>for (t = 0 : T: N^2 - T \ (T \ll N^2))</td>
</tr>
<tr>
<td>1. (Z = Z_t(n, t+1:t+P)).</td>
</tr>
<tr>
<td>2. Compute (B) by</td>
</tr>
<tr>
<td>(B_{\theta^{opt}}(i,j) = \exp \left( \sum_{p=-P}^{P} \alpha_{p} Q^{\ell}(i + p, j + p) \right))</td>
</tr>
<tr>
<td>(B = [(B_{\theta^{opt}}^1)^T, (B_{\theta^{opt}}^2)^T, \ldots, (B_{\theta^{opt}}^L)^T]^T)</td>
</tr>
<tr>
<td>3. Compute the new representation of the unlabeled data (4.12)</td>
</tr>
<tr>
<td>(\tilde{U} = \text{diag}( (1B_{\theta^{opt}}^{-1} ) B_{\theta^{opt}}^T U) ).</td>
</tr>
<tr>
<td>4. Compute the likelihood ratio for a new unlabeled frame</td>
</tr>
<tr>
<td>(\Gamma_t = \frac{f(\tilde{U}(t,:); H_1)}{f(\tilde{U}(t,:); H_0)}).</td>
</tr>
<tr>
<td>5. The decision rule for an unlabeled time frame is given</td>
</tr>
<tr>
<td>(H_1)</td>
</tr>
<tr>
<td>(VA_t = \sum_{j=-J}^{J} \Gamma_{t+j} \geq T_h).</td>
</tr>
<tr>
<td>(H_0)</td>
</tr>
<tr>
<td>6. Use (VA_t) to obtain the final VAD decision.</td>
</tr>
<tr>
<td>end</td>
</tr>
</tbody>
</table>
Chapter 5

Conclusion

5.1 Summary

In this thesis, we have addressed the problem of speech enhancement systems in presence of environmental noise, and presented two methods for overcoming this problem. The first, a supervised learning algorithm for voice activity detection based on representation of speech signals by spectral clustering, i.e., labeling, separating speech and non speech frames. This algorithm is composed of learning and testing stages: A sample cloud is produced for every signal frame by utilizing a moving window. MFCCs are then calculated for every sample in the cloud in order to produce an MFCC matrix and subsequently a covariance matrix for every frame. Utilizing the covariance matrix, we calculate a similarity matrix using spectral clustering and diffusion kernels methods. Using the similarity matrix, we cluster the data and transform it to a new space where each point is labeled as speech or nonspeech. We then use a GMM in order to build a statistical model for labeling data as speech or nonspeech. The main challenge was providing good results in presence of environmental noise and transient noise in particular. The key features of the proposed algorithm are the covariance matrix calculations via sample clouds and the novel similarity matrix computations. We demonstrated better results compared to a work that has already been proven to be superior to conventional methods of dealing transient noises, especially in cases of low SNR and Simulation results demonstrate its advantages compared to a recent VAD algorithm [32].

The second method is a constant beamwidth beamformer. The proposed approach
utilizes custom-tailored FIR filters for each microphone channel, manipulating the beampatterns beamwidths. The manipulated beampatterns have a constant beamwidth over a range of frequencies in a wideband. Additionally, a post summing output normalization filter is used to ensure a frequency invariant gain of the beampattern. By exploiting the physical microphone array configuration and attributes, we shape accordingly the frequency response of the FIR filters and eventually control the beamformer beamwidth. The proposed approach demonstrates improved array response results in various scenarios, compared to available methods in the field [16,20,63], especially in terms of sensitivity to parameters mismatch, noise robustness and side-lobe attenuation.

First, an input FIR filter array is designed to obtain a fixed main-lobe width, in a wide frequency band. Then, an output FIR normalizing filter is designed to keep a fixed gain over that band. The proposed design method achieves an approximation to a constant beamwidth output response, with better side-lobe attenuation performance and robustness to array mismatches, as compared to other methods examined. The new method is based on designing the FIR filter array such that its frequency response reduces the number of what we call, "effective microphones", when signal frequency increases, affecting an approximate fixed beamwidth in a wide band of frequencies. Simulations of the proposed beamformer also demonstrate better side-lobe attenuation and sensitivity performance, especially at the lower frequency spectrum. Moreover, the proposed algorithm offers a small deviation from the desired beamwidth in a wide band as well as low computational complexity due to its simple closed form FIR filter design procedure. Additionally, we have shown that when inducing random microphone gain and location inaccuracies, the proposed algorithm produces steady beampatterns throughout the frequency spectrum.
5.2 Future Research

This thesis unveils numerous possible research directions that can be studied:

1. The two speech enhancement methods proposed in this research thesis, i.e., the VAD system and constant beamwidth beamformer were individually researched. Integrating these two systems, as shown in Fig. 1.1, could yield a speech enhancement system robust to environmental noises. The beamformer degree would filter signals from undesired directions, effectively attenuating interference, noise and signals other than the desired signal. The VAD degree would further enhance the input signal since it would determine a desired speech signal from an undesired environmental noise, even if they originate from the same direction. In addition, this method would enable better power efficiency thanks to the fact that it would operate only when a desired signal from a desired direction is present.

2. A source localization system could also be implemented by integrating these two methods. Source localization is necessary for applications as automatic beamformer steering for suppressing noise and reverberation. Directing the main-lobe to a desired direction could be done by adding appropriate delays or phase shifts. This ability is necessary especially in applications where not only the location of the source is needed but also the ability to know when the source was active, as in teleconferencing and automatic video tracking. The straight-forward design of the beamformer allows us to easily introduce the delays or phase shifts that are required for the beam steering.

3. Research of an algorithm that would find the optimal placing of microphones in space so that the beamformer performance specifications can be met. Most microphone array design algorithm dictate the placing of the array microphones and force a specific geometry, e.g., ULA and circular arrays. In many cases, this drags substantial handicaps on the beamformer capabilities and performance. For instance, ULAs can only handle one source direction at a time resulting in uncertainties and possibly direction ambiguities. Therefore, ULA configurations must be ruled out when handling several signal sources. A novel beamformer design method that is
based on optimization can be researched. This method would take into consid-
eration only the desired beamformer specification, e.g., the desired beamwidth or
frequency range of interest, and would derive the placing of the microphones to
achieve these desired specifications. Optimization of the microphone placings in
space is a nonlinear, non-convex optimization problem. Thus, simple optimization
methods would not give the best results. Genetic algorithms have been researched
to find solutions in nonlinear, non-convex scenarios where conventional methods do
not provide sufficient results [74]. Genetic algorithm research in the field of beam-
forming and microphone placing in particular was recently done [75]. Yet, it is a
small niche in the world of array processing and much additional contribution can
be made.
Bibliography


ตน וניתוח של מעצב אלומה קבועה בתדר חיבור על מחקר
לשם מילוי חלק של הדרישות של קבלה תואר שני במססר לתאר
בהנפקת שמעל

אורן רוזן

הוגש לסנט הטכניון - מכון טכנולוגי לישראל
ה'תשע"ו בחיפה פברואר 2016
תודות

המחיקה נעשתה בחנאות פרופ' יוראלא כהנaghan מהפקולטה להנדסת חשמל בטכניון.

ברצוני לה благодא על התמיכת-professor יוראלא כהנaghan על ההנחייתה,

הנחייה וההדרכה Çalışת עבורה המחבר

ברצוני לה благодא על התמיכת-professor דוד מלאך על עזרתו והמיצצת בתוכחת

עבורה

לבמה, ארבעה בן המחודד לאבי ולאשתי נשל, על המיצמה

וסבלנותם בכל שלבי עבורה.

עבודה זו נכתבה לזכרה של אמי, אילנה רוזן ז"ל.
שקטות ערובות אלわけではないים לשיפור ציוס ההkoneksi בין אות רצוי להספק אות המוגדר כפרעה במגוון מערכות עיבוד אות ותקשורת כדוגמת: מכ"מ, רסונא, עיבוד אותות רפואי, תקשורת, טלוויזיה, חישה סיסמית ועוד. אחד השימושים הקלאסיים לשיטות עיצוב אלומה הוא היכולת ליב起こו רצויים ולהנחתת לבצע סינון מרחבי ובכך לגרום להעברת אותות מכיוון אחר אחד של הפוסים בין בריחים, ביטול התדואדועים, איכון מקור ושימושים רבים נוספים.

ה bucון קלע של אלגוריתמי תקציף אלומה לדגימה נלקח דגימה עיבוד מיקום המיקורים, MVDR סטרוטגי, שארטר בações על פליקריצית האוץ, ומניח את האוץ רובוט וחום (Spatial processing) שחריר, גרסה teasing על האוץ רגישות ותא צרכנות הפוסים, לCAF מעבר של הצעות האוץ רובוט סטר טט של ת퍅 (Gain) שיקוף במספר מעבדות באיכות ואוצרת את ביצוע הטקסט. בפרטים, האוצרת מעצב האוץ רצוי ערב אלגוריתמי רובוט סטר בטיהור האוץ אופנים קבוצת עבר סטר בטיהור התדרים, רחוב, קר, או שיפורים רציונליים ועבר של אזורי עיבוד蜉ח התדרים והאמת.

כימור בתוכנית מססר מזג פנים עכים מעבר מעצב אלומה קבוע בדרה. עיבו הצל מצרית אל קוודר יידי הביצועים הא쮮ים במקורי אוצרת רמה רמת בצירוא- ebay, הנחתת הפרשות והעבשות יריעה של קול רגישות בפורמטים נاعتمים מסדר. בשתייהdemandhypothetica על כל מיקורים ומיקורות אזורים במערך, עיבוד מעבר מיקורי דרך מגבלות הדרישות של תחום התדרים שמקורות אזורים ניעטרים, חוכם מעצב אלומה קבוע בדרה גל רגישות נ⧼ ו瞭解並 מתמדים nationally מוסכמים.

 ainda, מתכנת מיקורי דרגה הביצועים המתח גבוה בגובה רמה זמין, הנחתת הפרשות והעבשות הדריש רגישות בפורמטים נاعتمים מסדר. בשתייהdemandhypothetica על כל מיקורים ומיקורות אזורים במערך, עיבוד מעבר מיקורי דרך מגבלות הדרישות של תחום התדרים שמקורות אזורים ניעטרים, חוכם מעצב אלומה קבוע בדרה גל רגישות נrów ו라도 ו апрונים נishments הנוחים.

 trabalho זה מציג שיטה חדשה בתוכנית מעצב אלומה קבוע בתדרים. שיטה זו מבצעת שימוש במסנני תגובת מקל פיניט (finite impulse response filter) במהרכבה אלומת נשלות במקורי המחשה המסכים ובמקורי המחשה המסכים. שיטה זו מציגה את השיטה מחוץ למקורי מעצב אלומה קבוע בתדרים. שיטה זו מחזיקה את המיקורוס מסכים ובמקורי המחשה המסכים ובמקורי המחשה המסכים. שיטה זו מחזיקה את המיקורוס המסכים ובמקורי המחשה המסכים. שיטה זו מחזיקה את המיקורוס המסכים ובמקורי המחשה המסכים. שיטה זו מחזיקה את המיקורוס המסכים ובמקורי המחשה המסכים.
The steerer motion can be controlled using (beam steering) the concepts of wavefronts, and it can be directed to the right or left. The steerer motion is applied using various systems such as advanced tracking systems and other systems. A secondary method to improve the performance of the beam orientation is the integration of a speech recognition system. Speech recognition is a process of identifying speech segments and noise segments. A wide use of speech recognition is found in various applications such as speech recognition, speech synthesis, speech coding, and speech communication.

We conducted a study in a room that has a stationary noise environment and showed that even in transient noise environments, the recognition system can work. Transient noise is a short time and high energy signal, such as keyboard and door slams. These methods identify these noises and reject speech, not as a disturbance. The main reason for the false identification of these systems is the assumption that the noise changes slowly compared to speech.

In order to deal with the transient noise problem, and develop an algorithm for speech recognition that is based on spectral clustering (spectral clustering) and diffusion kernels (diffusion kernels). The algorithm is based on reducing the dimensionality of the representation of the speech using property vectors, which we refer to as a sampling cloud. After that, we use this information to compute a statistical model for the noise and speech information using a Gaussian Mixture Model (Gaussian Mixture Model). It shows a significant improvement in the performance of the proposed algorithm compared to the latest algorithms in the noise environment. In this way, we can communicate between listeners and improve the performance of the speech recognizer even when the recognizer is not confident.

In conclusion, this study and experimental results show that the proposed algorithm can help improve the performance of speech recognition systems and can be useful in various applications.