IMPROVED MICROPHONE ARRAY DESIGN WITH STATISTICAL SPEAKER IDENTIFICATION METHODS

KADIR ERDEM DEMIR
B.S., Electrical Engineering, YILDIZ TECHNICAL UNIVERSITY, 2009

Submitted to the Graduate School of Science and Engineering
in partial fulfillment of the requirements for the degree of
Master of Science
in
Computer Engineering

IŞIK UNIVERSITY
2016

IŞIK UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

IMPROVED MICROPHONE ARRAY DESIGN WITH STATISTICAL SPEAKER IDENTIFICATION METHODS

Kadir Erdem Demir

APPROVED BY:

Assoc. Prof. M. Taner Eskil (Thesis Supervisor)	Işık University	_____
Prof. Dr. Mustafa Karaman (Thesis Co-Advisor)	Istanbul Technical University	_____
Prof. Dr. Ercan Solak	Işık University	_____
Prof. Dr. Ahmet Aksen	Işık University	_____
Assoc. Prof. Onur Kaya	Işık University	_____

APPROVAL DATE: ..../..../....

IMPROVED MICROPHONE ARRAY DESIGN WITH STATISTICAL SPEAKER IDENTIFICATION METHODS

Abstract

Conventional microphone array implementations aim to lock onto a source with given location and if required, tracking it. This implementation is straightforward when the location or the path of the source and interference are provided. It becomes a challenge to detect the intended source when multiple unknown sources exist in the same environment.

Performance of speaker identification degrades drastically when the speech signal is severely distorted by additive noise and reverberation. In such environments, microphone arrays are often utilized as a means of improving the quality of captured speech signals.

Both microphone array and speaker identification are mature fields. The advances of these two distinct fields can be combined into one system that maximizes gain on the intended speaker, which is the topic of this thesis. We utilize microphone array methods to improve the accuracy of speaker identification in a cocktail party environment. When the source and interferences are localized microphone array can be tuned to further reduce noise and increase the gain.

In this thesis we developed a robust simulation environment to demonstrate the proposed improved microphone array design with statistical speaker identification. This is an open source implementation in which users can assign speakers anywhere in the room. We proposed two features; fusion based, and computationally efficient *N-Gram* for speaker identification. We demonstrated that the proposed features and the algorithm that leverages the synergy of microphone array processing and speaker identification methods outperforms conventional algorithms.

İSTATİKSEL SES TANIMA METODLARI İLE GELİŞMİŞ MİKROFON DİZİSİ TASARIMI

Özet

Mikrofon dizilerinin kazancı dizinin boyutlarını büyütürek artırılabilir fakat kazancı artırmak için sensör eklemek çok maliyetlidir. Bu nedenle eğer ortamda yeterince alan olsa bile algoritma karışıklığını artırarak kazancı artırma tercih edilir. Spektral dizi işleme methodlarında, odaklanılmak istenen kişinin ve gürültünün bulunduğu posizyonların bilinmesi büyük avantaj sağlar. Geleneksel metodlar bu problemi istatiksel olmayan yöntemlerle çözmeye çalışır. Ayrıca ses tanıma metodlarının performansları gürültü oranın yüksek olduğu ortamlarda azalır. Bu gibi ortamlarda, mikrofon dizilerinin kullanılması ses sinyalinin kalitesini artırır. Bu nedenlerde dolayı, mikrofon dizileri ve ses tanıma metodları birbirlerine katkı sağlarlar.

Bu çalışmamızda, mikrofon dizisi sistemi ve ses tanıma sistemi tek bir sistemin parçaları olarak tasarlanmıştır. Mikrofon dizisi kullanarak ses tanıma sisteminin doğruluğu artılırken ses tanıma sisteminin sonuçları kullanılarakta mikrofon dizisinin kazancı artırılmıştır. Ses tanıma sistemi uygulumasında *Fusion* ve *N-Gram temel frekans* yöntemleri önerilmiştir Gelişmiş mikrofon tasarımını gösterebilmek için simulasyon ortamı konuşmacıların odanın herhangi bir yerine eklenebiliceği bir simulasyon ortamı geliştirilmiştir. Simulasyon ortamında deneyler sonuçu önerilen metodların geleneksel metodlar üstün olduğu gözlemlenmiştir.

Acknowledgements

There are many people who helped me make my years at the graduate school most valuable. First, I would like to express my deepest acknowledgements to my supervisor Assoc. Prof. M. Taner Eskil and my co-adviser Prof. Mustafa Karaman for their support and guidance throughout my thesis work. I am also thankful to Işık University for providing me all the facilities and for being a part of such a wonderful environment. Finally, I will always be indebted to my wife Nataliya and my family for their patience and encouragement.

To my family . . .

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANN	**A**rtificial **N**eural **N**etworks
DFT	**D**iscrite **F**ourier **T**ransform
DSP	**D**igital **S**ignal **P**rocessing
DSR	**D**istance **S**peech **R**ecognition
FFT	**F**ast **F**ourier **T**ransform
FIR	**F**inite **I**mpulse **R**esponse
GMM	**G**aussian **M**ixture **M**odel
HMM	**H**idden **M**arkov **M**odel
IDFT	**I**nverse **D**iscrite **F**ourier **T**ransform
IIR	**I**nfinite **I**mpulse **R**esponse
LP	**L**inear **P**rediction
MFCC	**M**el-Frequency **C**epstral **C**oefficients
SVM	**S**upport **V**ector **M**achine
TDOA	**T**ime **D**ifference Of **A**rrival
VQ	**V**ector **Q**uantization

Chapter 1

Introduction

Microphone arrays are composed of multiple sensors. Each sensor in a microphone array functions as a single directional input device. Sound sources can be spatially filtered or distinguished from each other using the principles of sound propagation. By combining individual sensors, signals can be distinguished based on their spatial locations. In other words aperture signals are combined in a phased array in such a way that signals at particular spatial position experience constructive interference while others experience destructive interference allowing spatial filtering of a signal. This procedure is known as "beamforming" or "spatial filtering" [1].

Figure 1.1: A microphone array

Microphone array processing is a mature field with many uses. In corporate conferencing systems microphone arrays are widely used. An array on a conference room table or mounted on ceiling along with state of the art steering algorithms,

Figure 1.2: Flow Diagram of this thesis

digital signal processing, and echo cancellation, can zoom in acoustically on individual conference participants and deliver sound quality that is superior to traditional conference room sound-gathering methodologies.

Distance Speech Recognition (DSR) systems use microphone arrays and have useful applications in intelligent home and office environments. For instance by recognizing a speaker's voice even from long distance with the help of microphone arrays, an application can turn on a particular source of light in a room. The application makes it possible to have the light turned on just by saying: "Turn on the light.".

Microphone array processing also has medical applications such as hearing aids. The array provides significant improvement in speech perception over existing hearing aid designs, particularly in the presence of background noise, reverberation, and feedback [2].

Last but not least forensics is an important field of microphone array processing. With spatial filtering properties of microphone array a person can be overheard from a long distance without being observed. The results of overhearing conversations by using microphone array processing can be used as evidence in court according to law to prove whether a defendant is guilty or not.

Having voice samples of a person to be overheard and a microphone array we can locate the target through spatial localization and speaker identification techniques. If the speaker claims to be of a certain identity and the voice is used to verify this claim, this is called speaker verification. Combining microphone array with speaker verification promises strong contributions on forensics.

In this thesis we worked on two well known fields: microphone array processing and speaker identification. By using state of art methods these two disciplines will be combined into one system. The idea behind combining these two fields is to provide better speaker identification results using Microphone Array Processing techniques. Moreover, speaker identification distinguishes between the speaker and noise, which in turn will be used to tune microphone array gain. Thus the goal of this thesis is to increase the Microphone Array performance by using speaker identification techniques.

In this thesis a combination of two mostly used separate features, MFCC and Formants will be examined. A faster model for processing features "F0 N-Grams" model will be proposed and compared with conventional GMM models. As will be shown in chapter II microphone array gain increases in high frequencies. We will take advantage of that microphone array feature and focus on speaker recognition with using features on high frequencies.

Organization of this thesis is summarized in 1.2. In chapter II the first step of this chain process; spatial filtering with conventional beamforming methods will be examined. In chapter III state of art methods on speaker identification will be examined. First of all the applications of speaker identification will be introduced. Then the features used in identification and models used for processing these feature vectors will be introduced. In chapter IV post filtering microphone array processes will be introduced. Chapter V will introduce our speaker identification results. Chapter VI will be devoted to our approach for speaker localization with identification and post microphone array filtering processes. And Chapter VII will conclude this thesis.

Chapter 2

Conventional Microphone Array Processing

2.1 Introduction

Speech signals captured by a single microphone can be corrupted by undesired noise. As first step in this thesis we will examine conventional microphone array methods to overcome noise in this chapter. The position of the conventional beamforming processes can be seen in the picture 2.1.

Figure 2.1: Position of the conventional beamforming in project flow

Let's suppose the goal is to record one specific person in a crowded room and our target is located perpendicular to the recording device. If we use a microphone with an omnidirectional direction sensitivity our record will include other people's voices. But if we use several microphones instead of one and sum their outputs, total response of the system will be sensitive to direction. Summing process illustrated in Figure 2.2 for a microphone array with three elements.

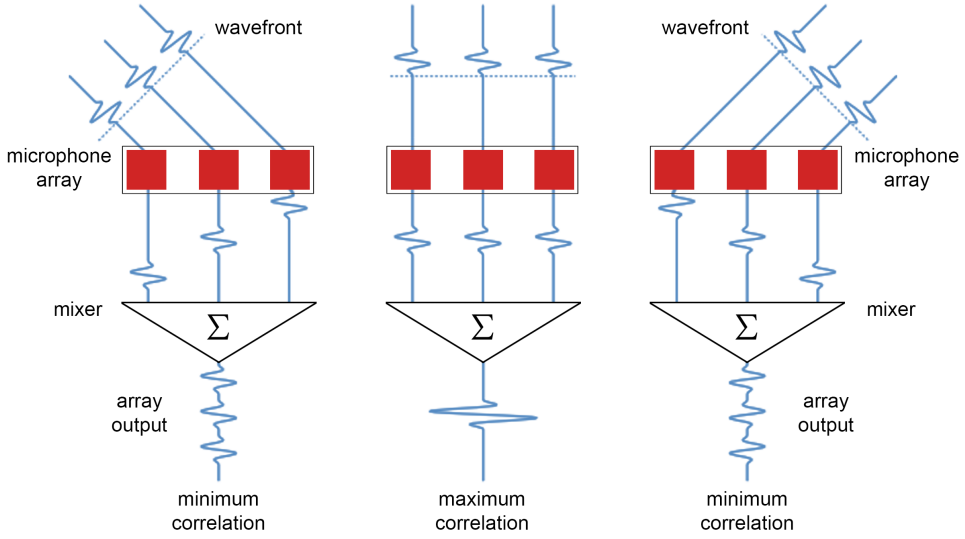


Figure 2.2: Linear microphone array setup

Because the array's output is created by summing all microphone signals, the maximum output amplitude is achieved when the signal originates from a source located perpendicular to the array; the signals arrive at the same time, they are highly correlated in time and reinforce each other. Alternatively, if the source originates from a non-perpendicular direction, the signals arrive at different time, they are less correlated and produce a lower output amplitude.

Even if signal processing techniques are not used, the amount of signal seen by microphone array varies with direction of microphone array. As illustrated in Figure 2.3 response of a microphone array is inherently directional. Response of microphone array as a function of frequency and direction of arrival (DOA) is known as *directivity pattern*.

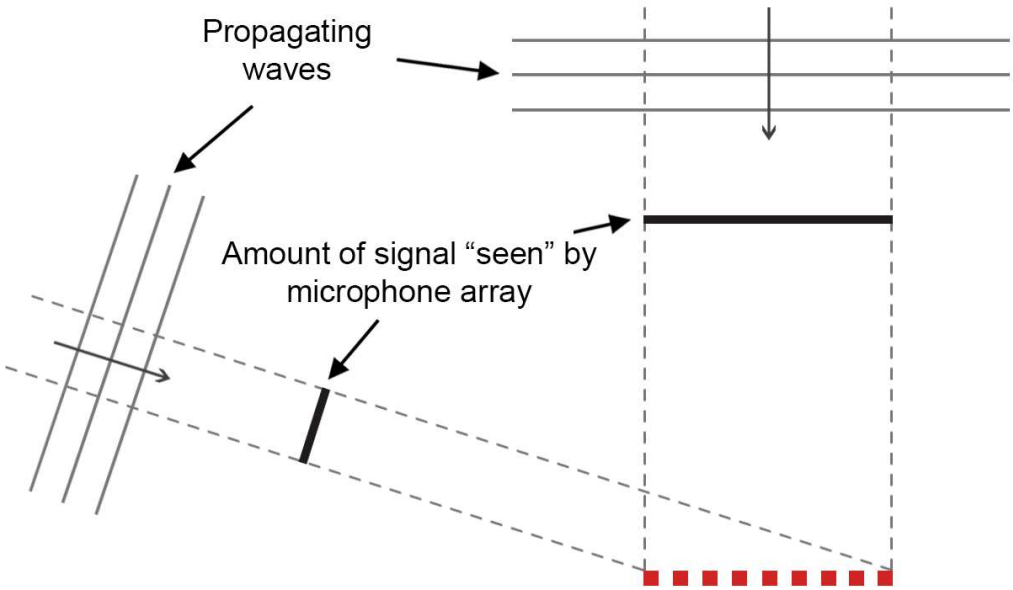


Figure 2.3: Directivity of a linear array

In this chapter mathematical model will be introduced for directivity pattern. This model will allow us to derive the parameters that affect the directivity pattern.

2.2 Description Of Problem

2.2.1 Mathematical Description Of Problem

In sensor arrays, due to the arrangement of the microphone array each element introduces some delay. The impulse response of each aperture is given by [3];

$$h_n(t, r) = \delta(t + \tau_n, r) \tag{2.1}$$

Where $\delta(t)$ is the dirac delta function. Additionally each element in a microphone array may introduce amplitude and time shift which is described by $w(t)$

in equation 4.2 ;

$$h_n(t,r) = w(t,r)\delta(t+\tau_n, r) \tag{2.2}$$

A signal can be reconstructed spatially by sampling the signal at a given instant of time. Frequency response can be derived by applying the Fourier transform [3];

$$H_n(f,\alpha_x) = W(f,r)e^{j2\pi\alpha_n r} \tag{2.3}$$

The α_n term in equation 2.3 is coming from k term of the sound wave equation. Wavenumber k is equal to $2\pi/\lambda$ and r is the spatial position of the aperture.

Given that microphone array has N elements, the output of array will be the superposition of all elements given by [3] ;

$$D(f,\alpha_x) = \sum_{n=-N/2}^{N/2} W(f,r)e^{j2\pi\alpha_n r} \tag{2.4}$$

In this chapter we will consider elements equally spaced. And elements themselves does not introduce any amplitude gains or phase shifts. So element weight $W(f,r)$ will be taken as 1. Directivity pattern will be equal to 2.5 ;

$$D(f,\theta) = \sum_{n=-N/2}^{N/2} e^{j\frac{2\pi f}{c}nd\cos(\theta)} \tag{2.5}$$

The Equation 2.5 is well known Fourier transformation of a rectangular window. And result of such transformation is given [3];

$$D(f,\theta) = L\frac{sin(x)}{x} \tag{2.6}$$

According to Eq. 2.5 similarities between Time - Frequency domains and Sensitivity - Directivity domains can be derived. As we can see directivity response of microphone array is like a Finite Impulse Response(FIR) filter. Each element in array corresponds to one weight of FIR filter window. And element count corresponds to window size. [4].

From equation 2.5 it can be seen that performance of a microphone array depends on frequency, number of elements and distance between microphones. Each of these parameters will be tested in our simulation environment.

2.3 Spatial Aliasing

In temporal sampling, nyquist frequency states the minimum rate at which a signal can be sampled without introducing errors, is twice the highest frequency present. In spatial filters spatial aliasing can occur as well.Aliasing can happen in space, as well as in time. In order to reconstruct a spatial sinusoid from a set of discrete samples, spatial sampling must occur at a rate greater than a half of the wavelength of the sinusoid. The relation between distance between elements and frequency is given by; [5]

$$d < \frac{f}{2c} \tag{2.7}$$

2.4 Near Field Behaviour

Until this point we considered plane waves but by its nature sound waves are spherical. As illustrated in Figure 2.4 the further the waves from the microphone array more straight they appear. The region where waves appear as a plane wave is called far-field.

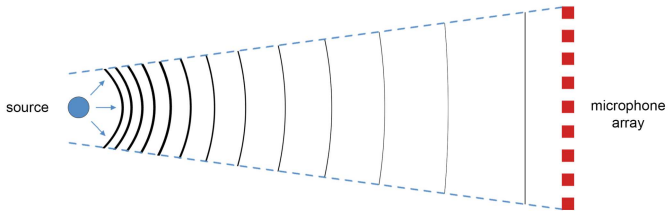


Figure 2.4: Further the source from the microphone array, waves become more planar

For a linear array, a wave source may be considered to come from the far-field of the microphone array if

$$d > \frac{2L^2 f}{c} \tag{2.8}$$

2.5 Delay-sum Beamformer

So far, directivity pattern has been fixed perpendicular to the array. With beamforming capabilities of microphone array, beam pattern can be steered without moving array. And also microphone array can focus to different depths.

As illustrated in Figure 2.5 sound arrives first at the centre of the array and a few microseconds later at the outer edges. By delaying the sound at the centre until it is in phase with that coming from the outer edges, a more tightly focused beam can be generated. [6]

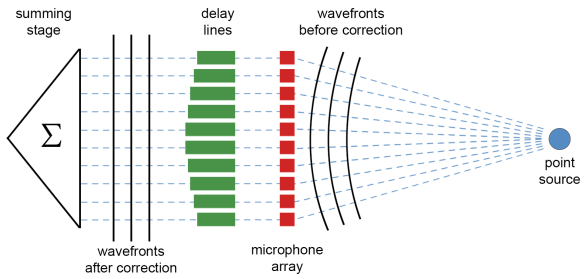


Figure 2.5: Focusing in depth with microphone array

Steering can be simply achieved by adding a delay stage to each of the array elements as illustrated in Figure 2.6 [7]

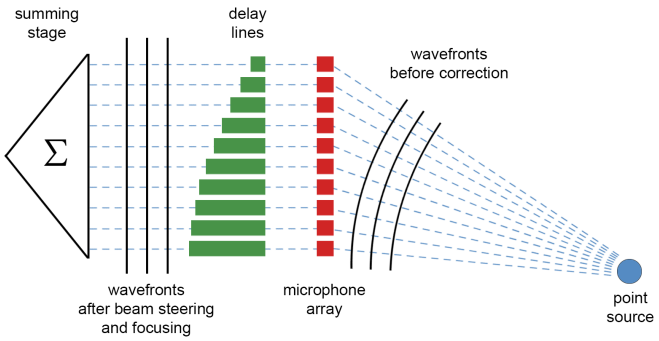


Figure 2.6: Steering microphone array

Delay-sum microphone array's response to a pulse at $0,33\pi$ is simulated and can be observed in Figure 2.7

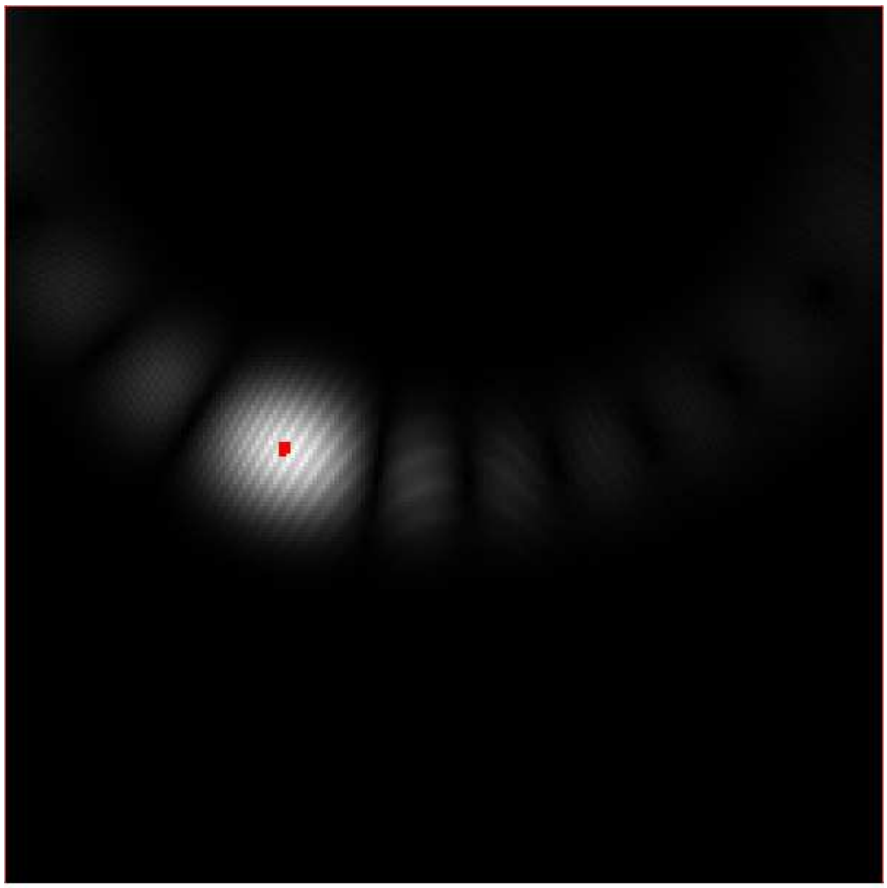


Figure 2.7: 2 dimentional spatial response of steered microphone array

2.6 Conventional Microphone Array Summary

We observed that delay-sum microphone array's behaviour is similar to a moving average FIR filter. It is important to note that conventional microphone array methods do not need any speaker or location data. These deterministic methods will help us to locate the target speaker and interference. After interference is filtered spatially we will see that speaker identification methods will provide better results.

Chapter 3

Speaker Identification

3.1 Introduction

The task of speaker identification, also referred as detection, is to determine speaker by a segment of speech. Generally segment of speech contains only one speaker and task is better termed single-speaker identification [8]. If speech segment contains more than one person than task becomes multispeaker detection. We will focus on single-speaker detection task on this chapter. But in this thesis we will also focus on multispeaker detection.

Larynx sizes, vocal tract shapes and other parts of voice production organs of individuals are different. Also each speaker has different manner of speaking, accent, rhythm, pronunciation pattern. In this chapter we will give an overview of features produced by these differences. Then we will give an overview to classical methods which use these features. [9].

Speaker identification is a part of this thesis and stand-alone modular entity. Identification process will start after conventional beamforming processing. The position of the identification process in project flow is illustrated in Figure 3.1. Since identification is a stand alone modular entity its applications and results will be over-viewed separately.

Figure 3.1: Position of the speaker recognition in project flow

One important field of speaker identification is telephone-base services. An example is automatic password reset over the telephone. Demand to such services are high and assigning human-operators for those kind of tasks are expensive.

Also voice identification is the most preferred form of biometric identification among consumers. With all types of biometric applications on the rise, voice-based authentication is one approach that seems to engender less resistance among users than other biometric forms of security.[10]

In this thesis we will use speaker identification for spatially locating the target speaker. It is particularly important to locate speaker location because given location information we can apply post filtering microphone array processing methods. The process of locating the spatial position of the speaker is illustrated in Figure 3.2. By using delay-sum beamformer's steering and focusing capabilities spatial filtering will be applied to each point of the room. And spatially filtered data will be passed to identification system. Since spatial filtering increases the gain of the focused point and eliminate the noises from other points when focused point is closer to speaker position our identification system will score better. The spatial position where identification system provides the highest score will considered as position of the speaker.

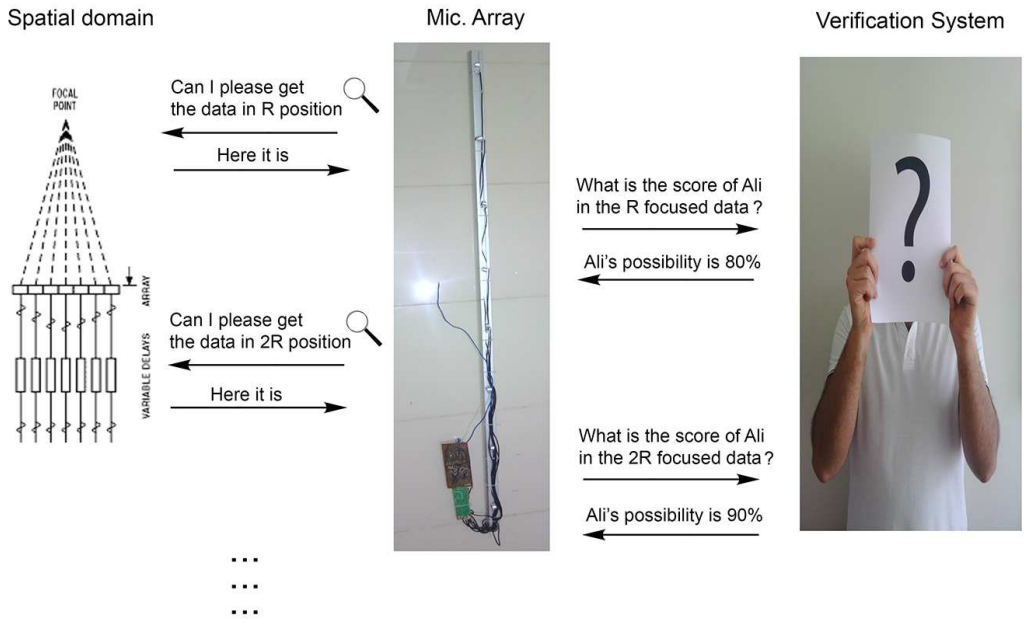


Figure 3.2: Spatially locating the speaker by combining delay-sum beamformer and speaker identification methods

Speaker identification systems can be divided into text-dependent and text-independent ones. In case of text-dependent systems, phrases known beforehand. Smaller the dictionary more accurate identification process will be. Text-independent systems are more challenging because there are no limitation on the words that the speakers are allowed to use. When training utterances and test utterances are different accuracy of speaker identification system might decrease. This is called phonetic variability and one of the biggest challenges of text-independent systems.

There are other factors which effect the performance of identification system. Changes in recording devices and the sample rate are a few examples. Another challenging problem in speaker identification is "Session variability" which is any variation between two recordings of same speaker. There might be many reasons of "Session Variability" such as aging, environment, health and mood [11].

In this chapter we will give an overview of features used in identification and models which uses these features.

3.2 Speaker Identification Fundamentals

A speaker identification system includes two primary components: a front-end and a back-end. Front-end component extracts features. Digital signal processing techniques like filter banks, fast fourier transform used in front-end. And also every speaker identification system has back-end component where speakers are modelled (enrolled) and identification trials are scored. Mostly machine learning and statistical tecniques are used in back-end.[12].

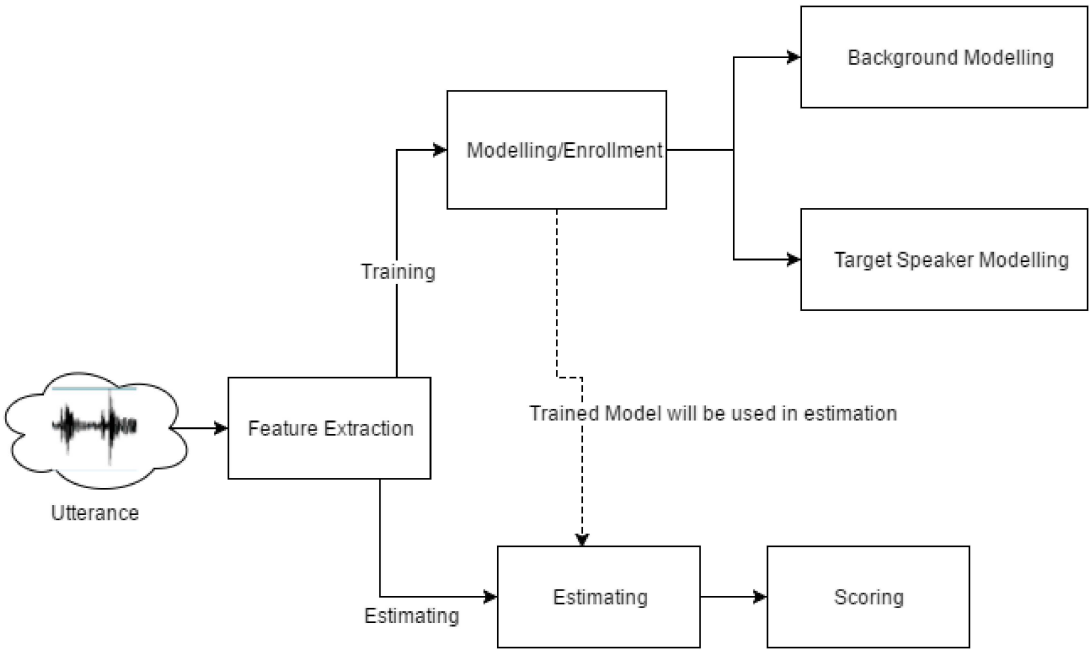


Figure 3.3: Stages of a typical automatic speaker identification system.

Figure 3.3 shows two stages of an automatic speaker identification system. The upper is the training process , while the lower panel illustrates the prediction process. The feature extraction module first transforms the raw signal into feature vectors in which speaker specific properties are emphasized and statistical redundancies suppressed. In the training mode; a speaker model is trained using the

feature vectors of the given speaker. While training some systems also use "anti-speaker" techniques, such as cohort models, or well designed world models which contains balance of voices that would be representative of the voices of potential impostors [13]. In the recognition mode, the feature vectors will be extracted from unknown utterances. And those feature vectors will be compared against trained model. Thus a similarity score will be get from models. Finally decision module will use this similarity score to make the final decision. In the recognition phase, background speakers are used in the normalization of the speaker match score. This simple approach can be seen in equation 3.1.

$$normalized\ score = \frac{claimed\ identity}{world\ model} \tag{3.1}$$

In upcoming sections we will give a introduction to features and modeling respectively. Than we will explain the methods used in this thesis in details.

3.2.1 Front-End Overview

In speech signal there are many features and some are not useful for speaker identification. The characteristics of and ideal feature would; [9]

- be independent recording environment,
- has as small session variability as possible,
- be robust against noise and distortion
- occur frequently and naturally in speech
- to be calculated fast enough from speech signal
- be difficult to impersonate/mimic

According to application importance of properties may change. For example in a real time application speed might be more important. But for a application which

is designed for noisy environments, small session variability will be preferred. We should not that feature selection not only effects feature extraction speed but also effects modeling process which uses this feature. Traditional statistical models such as the Gaussian mixture model cannot handle high-dimensional data fast enough. The number of required training samples for reliable density estimation grows exponentially with the number of features. This problem is known as the curse of dimensionality.[14]

Features can be divided into spectral features, temporal features, prosodic features and high-level features. The spectral features, which are obtained by converting the time based signal into the frequency domain using the Fourier Transform. Some examples of spectral features are fundamental frequency, frequency components, spectral centroid, spectral flux, spectral density, spectral roll-of are examples of spectral features. These features can be used to identify the notes, pitch, rhythm, and melody.

The temporal features are obtained in time domain. They are simple to extract and have easy physical interpretation. Energy of signal, zero crossing rate, maximum amplitude, minimum energy are examples of temporal features.

When sound put together in connected speech prosodic features can be observed such as intonation, stress and rhythm. Prosodic features span over tens or hundreds of milliseconds. [15]

For high-level feature extraction, input speech is converted into a series of tokens. The tokens are time-ordered discrete symbols and represent linguistically significant interpretations of the input signal. Examples of token types are words, phones, and pitch gestures. [16]

In next section features that will be used in this thesis will be explained in detail.

3.2.2 Spectral Features

To have stationary signal utterance must be broken down in short frames of about 20-30 milliseconds because speech signal continuously changes. Spectral envelope which is a curve in the frequency-amplitude plane, derived from a Fourier magnitude spectrum 3.4, contains information about the resonance properties of the vocal tract.

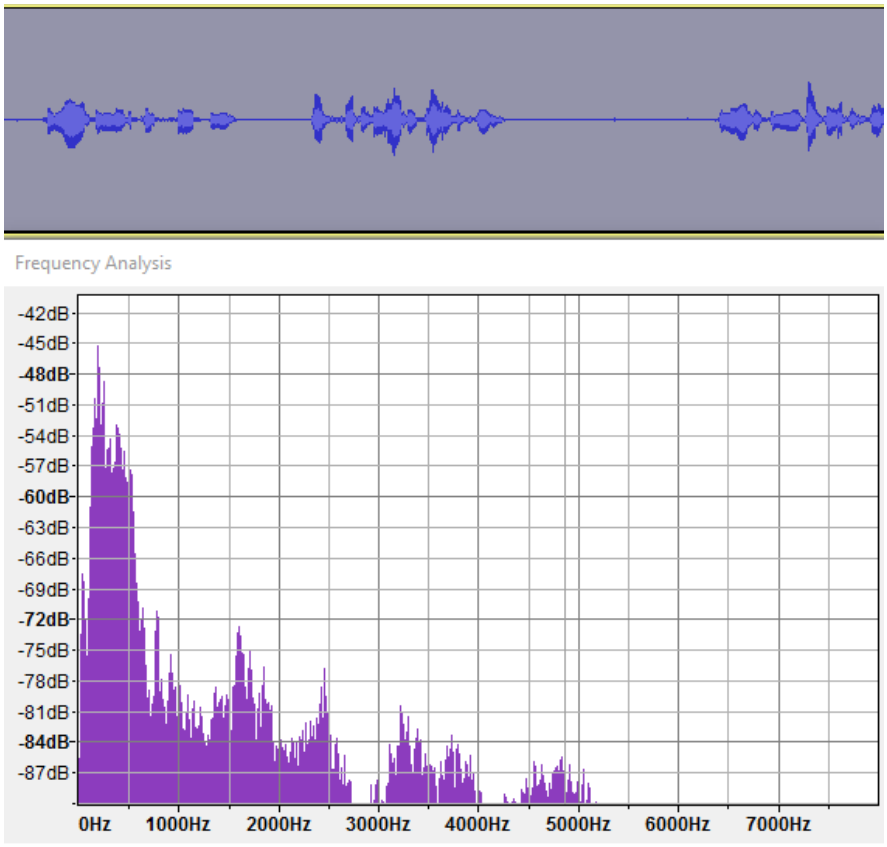


Figure 3.4: Typical magnitude spectrum of a human voice

Features can be extracted directly from spectral spectrum but usually using other transformations dimensionality is further reduced. Filter banks which are an array of band-pass filters that separates the input signal into multiple components, each one carrying a single frequency sub-band of the original signal are used for reducing dimensionality. Different types of filter banks have been proposed to embed the observations of psychoacoustic experiments about in the frequency analysis.

Mel-frequency cepstral coefficients (MFCCs) are popular features. MFCCs were introduced in early 1980s for speech verification. Mel filter banks applied to spectral spectrum for computing MFCC coefficient, followed by logarithmic compression and discrete cosine transform (DCT). Denoting the outputs of an M-channel filterbank as $Y_m, (m = 1, 2, ..., M)$, the MFCCs are given [17]

$$c_n = \sum_{m=1}^{M} [\log Y(m)] \cos[\frac{\pi n}{M}(m - \frac{1}{2})] \tag{3.2}$$

Here n is the index of the cepstral coefficient. M generally chosen between 20 and 40. In order to reduce dimensionality, final MFCC vector is obtained by retaining about 12-15 lowest DCT coefficients.

Gammatone coefficients and linear predictive cepstral coefficients(LPCC) are other important features calculated by filter banks.

3.2.3 Prosodic Features

Prosody refers to non-segmental aspects of speech, including for instance syllable stress, intonation patterns, speaking rate and rhythm. The most important prosodic parameter is the *fundamental frequency (or F0)*. F0 can be observed at the time domain representation of a human speech which illustrated in Figure 3.5. Each of the identifiable repeating patterns is called a cycle. The duration of each cycle is called the *glottal pulse* or *pitch period* length. The fundamental frequency of a periodic signal is the inverse of the pitch period length. The voiced speech of a typical adult male will have a fundamental frequency from 85 to 180 Hz, and that of a typical adult female from 165 to 255 Hz.[18]

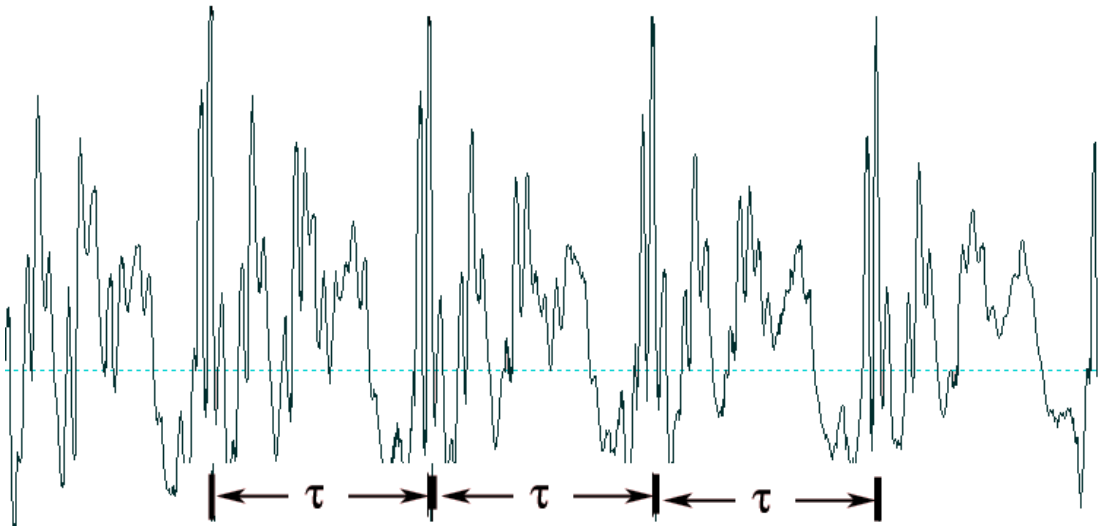


Figure 3.5: Fundemental frequency cycles in human speech

Reliable F0 determination itself is a challenging task because of problems illustrated in Figure 3.6. In 3.6.a simply selecting the highest peak will be enough but in case of 3.6.b F0's magnitude is less than its harmonics. In this case selecting the lowest frequency seems to work. When we check 3.6.c and 3.6.d we see that some peaks are missing. It is a challenge to accurately estimate F0 due to these issues.[19]

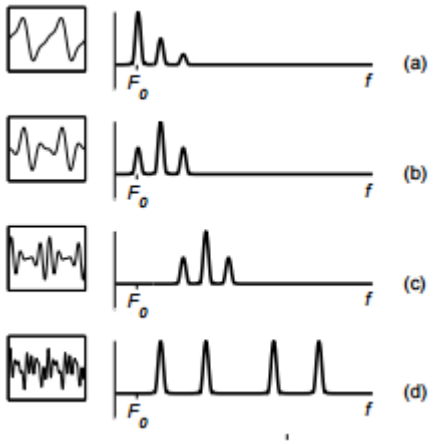


Figure 3.6: Fundemental frequency cycles in human speech

Formants are also considered as good parameters; a formant is a concentration of acoustic energy around a particular frequency in the speech wave. There are

several formants, each at a different frequency, roughly one in each 1000Hz band. Each formant corresponds to a resonance in the vocal tract. In Figure 3.7 first formant (F1) and second formant (F2) is shown by bold lines. [20].

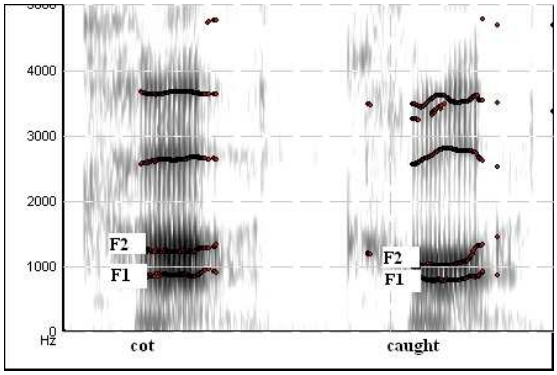


Figure 3.7: Formants can be seen very clearly in a wideband spectrogram,they are displayed as dark bands.

3.2.4 Speaker Modelling

After feature vectors are extracted, speaker model(s) are trained with extracted feature vectors. Classical speaker models can be divided into template models and stochastic models. In template models, training and test feature vectors are directly compared with each other and correlation between feature vectors represents their degree of similarity. Vector quantization (VQ) [21] is a representative example of template models.

In stochastic models, a probability density function assigned to each speaker. In training phase probability density function will be estimated. Estimation will be done by finding the probability of utterance in trained model. The Gaussian mixture model (GMM)[22] and the hidden Markov model (HMM) [23] are the most popular models for identification. We will examine the methods used in this thesis further.

3.2.5 Gaussian Mixture Model

Gaussian mixture model (GMM) is a stochastic model which is widely used in speaker identification. A feature vector belongs all class but with a certain ratio. In identification the speaker class which has the maximum likely-hood will be chosen.

For a D-dimensional feature vector x, the mixture density used for the likelihood function is given [8]

$$p(x|\lambda) = \sum_{i=1}^{M} w_i p_i(x) \tag{3.3}$$

The density is a weighted linear combination of M uni modal Gaussian densities, $p_i(x)$, each parametrized by a mean $D \times 1$ vector, $_i$, and a $D \times D$ covariance matrix,

$$p_i(x) = \frac{1}{(2\pi)^{D/2} \mid E_i \mid^{\frac{1}{2}}} (\exp -\frac{1}{2}(x - \mu_i)^{'}(E_i^{-1})(x - \mu_i)) \tag{3.4}$$

GMM is a method to represent high dimensional feature vector with a smooth surface of membership probability. This capability of GMM makes it superior to the Naive Bayes approach although it is computationally more costly.

3.2.6 N-Gram Model

In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sequence of text or speech. Items in our applications will be formants. In a speech signal formants change very frequently. And goal of our N-Gram tokenizer is create a model by observing this changes to detect speaker.

In training phase we extract F0 from different utterances of same speaker. We count occurrences of each frequency bin of F0. In GMM model each speaker have

her/his own probability density function, in N-Gram model each speaker will have a total number of occurrences for specific frequency bin. While estimating we extract F0 from test utterance. If C_n is the number of occurrences that extracted F0 bin occurs in trained model of speaker n, and T is total number of occurrences in all training models then for a 1-gram model probability of utterance belongs to speaker n is C_n/T. Each speaker's n-gram model will be tested. And model which gives the highest probability will be detected as speaker.

Using counts directly is not a good approach for speaker identification a single F0 frequency bin is not unique enough to distinguish people from each other. Instead of counting a single bin we offer a hash function in F0 normalization stage. This hash function creates a unique number out of N-bins. Than we count result of our hash function which composed of N frequency bins.

Chapter 4

Microphone Array Post Filtering

4.1 Introduction

In Chapter 2 it was told that microphone array element's may introduce amplitude and time shift which was described by $w(t)$ in Equation 4.2. Up to this point of discussion, we neglected amplitude gain and time shift introduced by microphone array elements and assumed equally weighted sensors calculating the directivity pattern, given

$$w_n(f) = \frac{1}{N} \tag{4.1}$$

In this chapter we will investigate effects of non uniform weight vector w. Firstly we will introduce narrow band and broad band beamformers and illustrate how they operate. As illustrated in Figure 4.1 we can reconstruct the signal over all space and time by either temporally sampling the signal at a given location in space, or spatially sampling the signal at a given instant of time[4]. To provide insight to different aspects of spatial filtering with a beamformer, FIR filtering methods will be used.

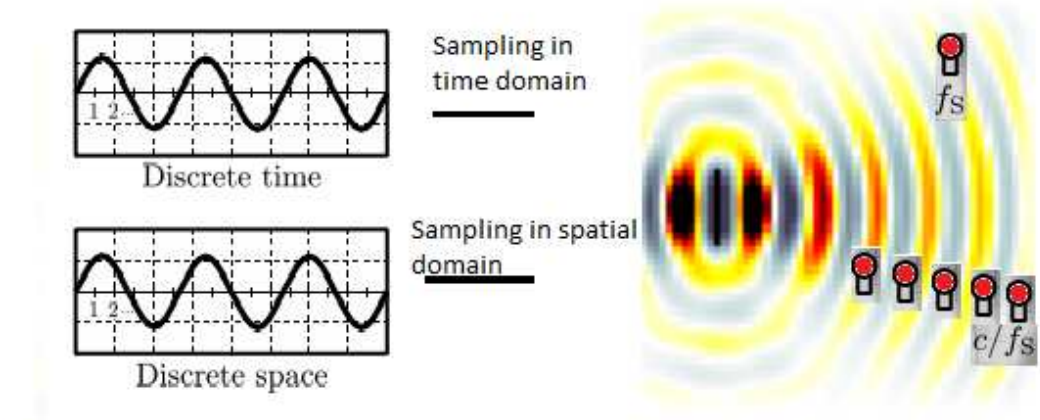
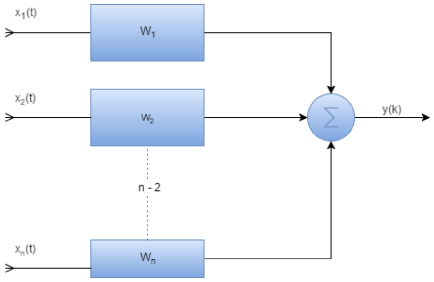


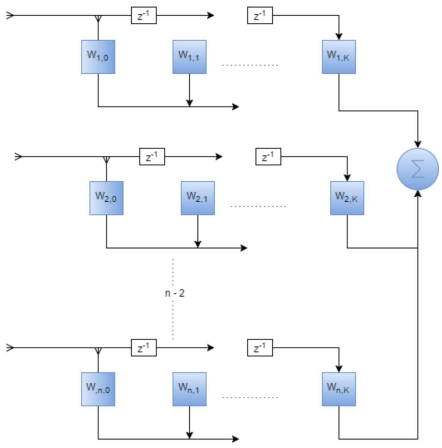
Figure 4.1: Reconstructing a signal with sampling by time and space

4.2 Narrow Band and Broad Band Beamformers

Two beamformers are illustrated in Figure 4.2. Beamformer depicted in Figure 4.2a is typically used for narrow band signals. The combination of sensor data is given by Equation 4.2.



(A) A common narrow band beamformer



(B) A common broad band beamformer

Figure 4.2: Illustration of narrow band and broad band beamformers, each sensor output is multiplied by a complex weight and then summed

The second beamformer which illustrated in Figure 4.2b samples both space and time and often used while filtering broadband signals. In this case output is given

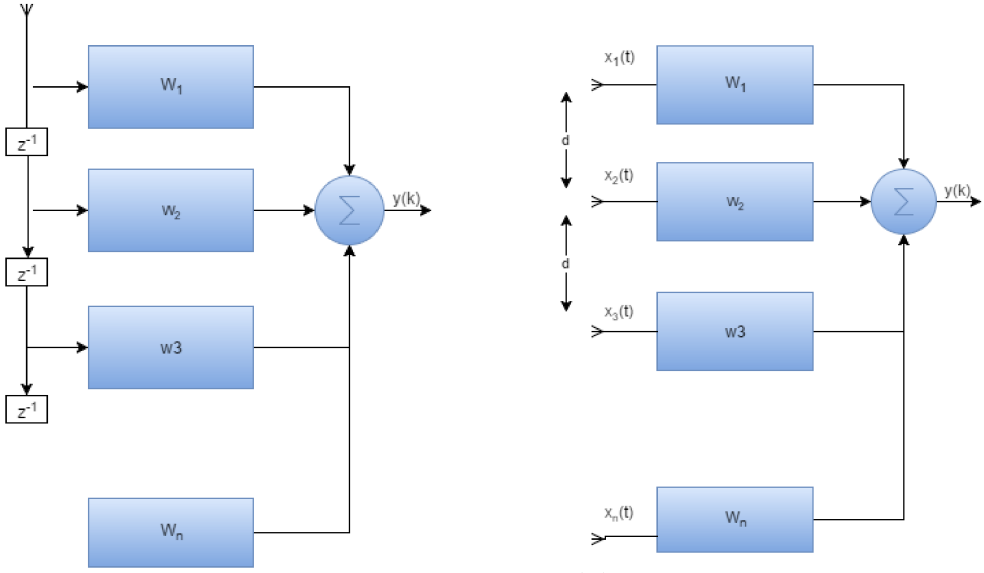
$$y(k) = \sum_{n=-N/2}^{N/2} \sum_{k=0}^{K} w_{j,k}(x(k - \tau_n)) \tag{4.2}$$

where N is number of elements in aperture and K is the number of delays in each element[24].

4.3 Similarities With FIR Filtering

Methods and techniques used in FIR filtering and spatial filtering are familiar because both use reconstructing signal as base. Although there are many similarities in someways beamforming differs from FIR filtering. For example, in beamforming source of energy has several parameters like range, elevation angle, polarization and temporal frequency. [24].

As illustrated in Fig 4.3 the correspondence between FIR filtering and beamforming is closest when the array geometry is linear and equi-spaced and the beamformer operates at a single temporal frequency ω_0.



(A) Equi-spaced omni directional narrow-band line array

(B) A single channel FIR Filter

Figure 4.3: Comparision between FIR filter and Equi-Spaced omni directional narrow-band array

In FIR filtering, a band-pass filter is a device that passes frequencies within a certain range and rejects frequencies outside that range. Likewise in spatial filtering we may wish to receive any signal arriving from a range of directions, in which case the desired response is unity over the entire range. As another example, in FIR filtering, band-stop filter is a filter that passes most frequencies unaltered, but attenuates those in a specific range to very low levels. Likewise in spatial filtering and we may know that there is strong source of interference arriving from a curtain range of directions, in which case the desired response is zero in this range. In spatial filtering instead of band-pass and band-stop terms, such filters are called $beamformer$[24].

4.4 Pattern Shaping

As described in 2 directivity of a microphone array is given by 4.3 ;

$$D(f, \alpha_x) = \sum_{n=-N/2}^{N/2} W(f, r) e^{j2\pi\alpha_n r} \tag{4.3}$$

Considering only the horizontal directivity pattern, we have

$$D(f, \theta) = \sum_{n=-N/2}^{N/2} W(f, r) e^{j\frac{2\pi}{\lambda} nd\cos\theta} \tag{4.4}$$

As described in previous chapters with speaker identification method we will locate the desired speaker and also the interference location. Given we know beam pattern we can figure out the weight vector of the aperture.

$$W(f, r) = \sum_{n=-N/2}^{N/2} D(f, \theta) e^{-j\frac{2\pi}{\lambda} nd\cos\theta} \tag{4.5}$$

To obtain weight vector, firstly beam pattern is obtained from equation,

$$D(f, \theta) = L\frac{sin(x)}{x} \tag{4.6}$$

Secondly nulls are placed in interference direction with a window function. And finally IDFT is applied to the windowed beam pattern to obtain weight vector.

Chapter 5

Speaker Identification Experiments And Results

5.1 Feature Selection And Modelling

In this study we used spectral features such as MFCC, formants and also prosodic features which are obtained from fundamental frequencies as illustrated in Figure 5.1.

In calculation of MFCC, the number of filter banks and selection of cepstrum coefficients are important. We used 40 filter banks and 39 cepstrum coefficients. The more cepstrum coefficients the better identification accuracy will be at the expense of computational cost in feature extraction and GMM modelling processes. To extract formants from speech signal we first calculated fundemental frequency by using Yin's FFT method [25]. As it was told in previous chapter there are several formants, each at a different frequency, roughly one in each 1000Hz band. Since formants are resonances in the vocal tract, formant frequency has a large amplitude. In each 1000Hz band we compared the frequencies which are multiples of F0 and selected the frequency with the largest amplitude. We used 5 formants as features. Both formants and MFCC were modeled using GMM with 5 clusters.

The prosodic characteristics of speech were modeled using F0 N-Gram model where N is 5. We also used F0 feature in a N-Gram model that N is selected as 3. We chose the highest frequency that avoids aliasing between 1000-2000Hz as was discussed in Chapter 2.

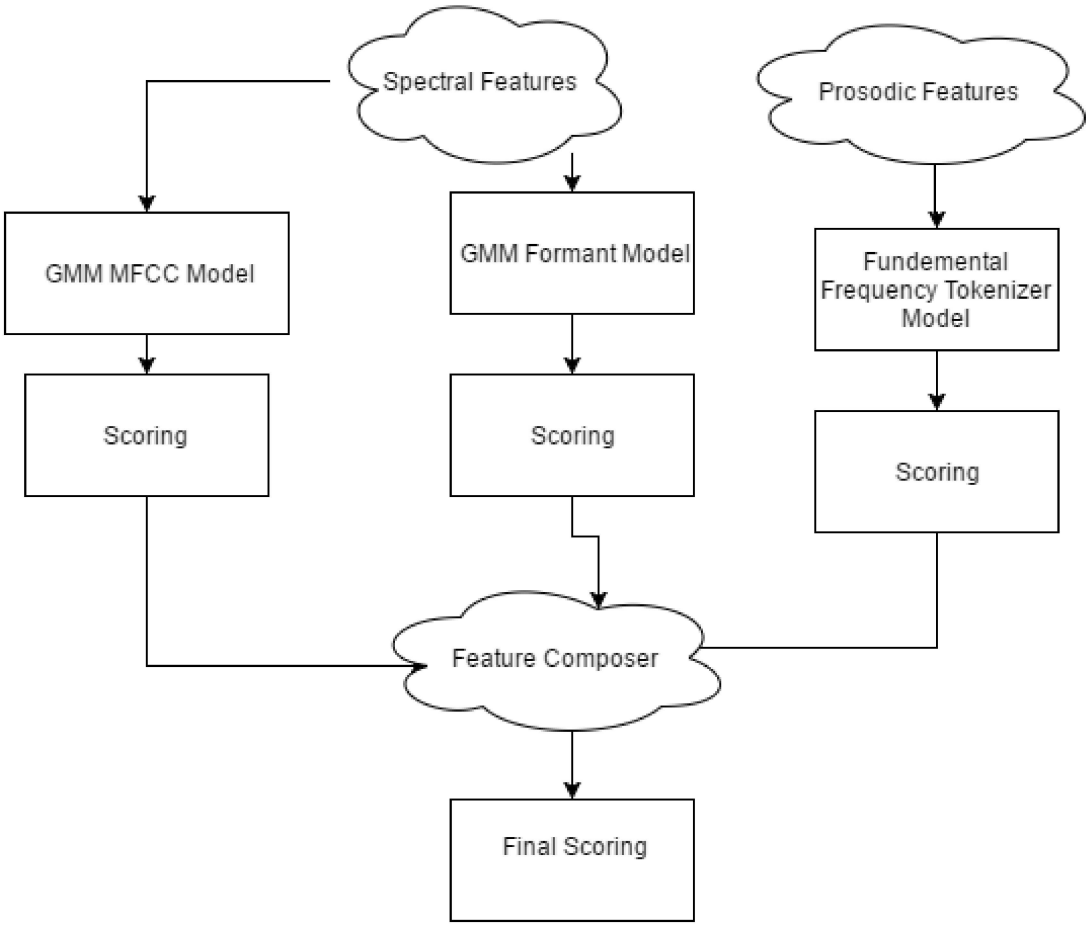


Figure 5.1: Illustration of methods used and scoring

5.2 Exprimental Results

For our purposes, detecting a speaker with the shortest possible sample is critical. We made experiments with varying duration of test utterances. Moreover since we are targeting real time, the speed of our algorithms is also a concern. In training and estimation stages we used EUSTACE [26] database which is composed of utterances collected from 10 people. In all of our experiments test and training data are mutually exclusive.

Accuracy of MFCC, formant and N-Gram F0 models with varying training data can be seen in the Table 5.1. Each feature models different aspects of speech. For example formants model subject's vocal tract shapes and N-Gram F0 targets

Train Data Length	MFCC	FORMANT	N-Gram F0	Fusion
30 Seconds	52.3492	45.2553	43.7391	65.0479
50 Seconds	77.5403	71.6232	55.609	87.8661
100 Seconds	79.8842	71.2382	56.1335	87.0293
125 Seconds	87.2329	73.5190	59.8617	99.4390
150 Seconds	91.2655	80.0374	60.2278	99.4048

Table 5.1: Percentage of correct identification with MFCC, Formant, NGram-F0 and Fusion Models where test data length is 500 milliseconds.

to model intonation, stress and rhythm. We also offer a combined model which is a fusion of features. Our motivation is to support our decision system when there are similarities between the sound, intonation, stress or rhythm between samples. For example vocal tract shapes of two individuals might be similar and formant modeling may fail to distinguish these two individuals. But if their intonations are different when we combine formant modeling and N-Gram F0 modeling, identification accuracy is expected to increase. The results of combined model can be seen in the fourth column of Table 5.1. In this expriment test data length was chosen as 500 milliseconds.

The accuracy of model with varying duration of test utterances can be seen in Table 5.2. In this experiment the training data was fixed at 50 seconds for each individual.

Test Data Length	MFCC	FORMANT	N-Gram F0	Fusion
200 mSeconds	69.7811	59.7131	48.4566	86.2620
500 mSeconds	77.5403	71.6232	55.6090	87.8661
800 mSeconds	81.4202	78.6349	58.8514	89.7959
1000 mSeconds	82.1940	81.2907	59.9640	92.0635
1500 mSeconds	83.9669	85.6510	62.5000	96.6667

Table 5.2: The effects of test utterance length in MFCC, Formant, NGram-F0 and Fusion Models where training data length is 50 seconds.

Algorithm speed comparison for training the models are shown in Table 5.3 with varying amount of training data.

Train Data Length	MFCC	Formants	F0 N-gram
10 Second	11.8000	3.0300	0.2600
100 Seconds	67.4080	19.2780	0.2800
250 Seconds	104.9080	30.3500	0.2900
500 Seconds	170.1780	66.1230	0.3150

Table 5.3: The comparison of training speed between 3 models. Result are written in seconds.

The proposed features and the results which are obtained in speaker identification experiments will be revisited in the conclusion chapter.

Chapter 6

Microphone Array Experiments and Results

6.1 Parameters of Directivity Pattern

6.1.1 Simulation Environment

We created a simulation environment of $20m^2$ to test the effects of frequency, number or sources,number of sensors and the distance between them. For simplicity the room is represented in two dimensions. A microphone array will be placed on the wall of the room. Sources are simple .wav files which can be placed anywhere in the simulation room. The number of microphones and distance between microphones are adjustable. Our simulation environment can be seen in Figure 6.1 below. In this example a microphone array with 41 elements and two sources are placed in the simulation room.

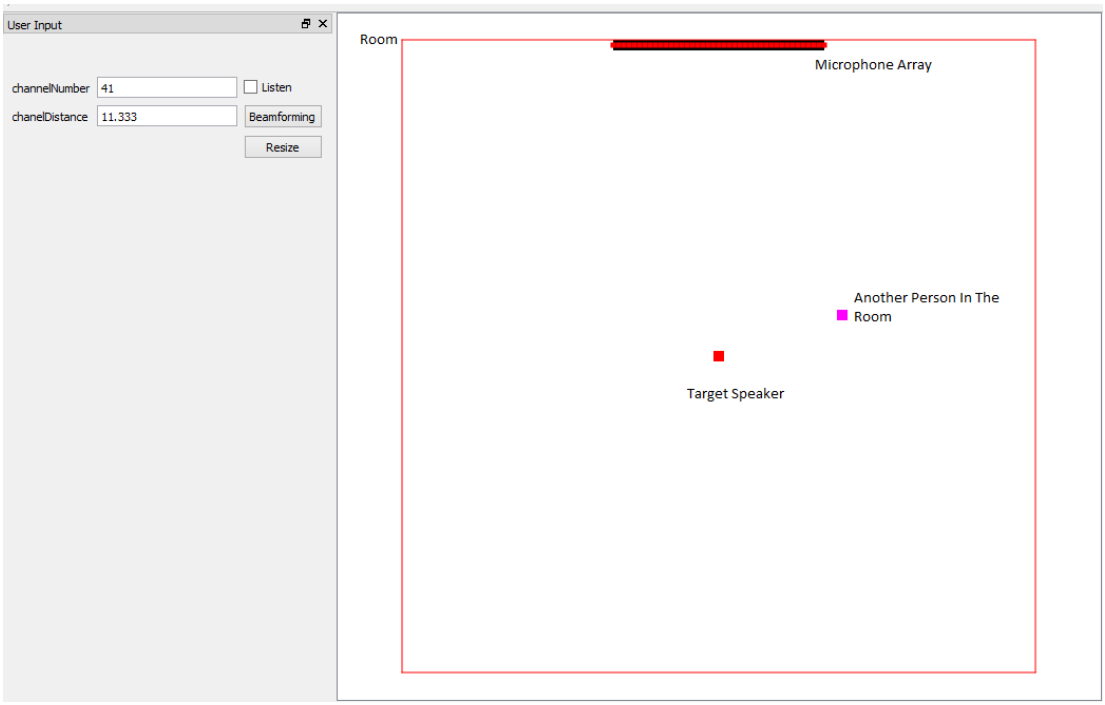
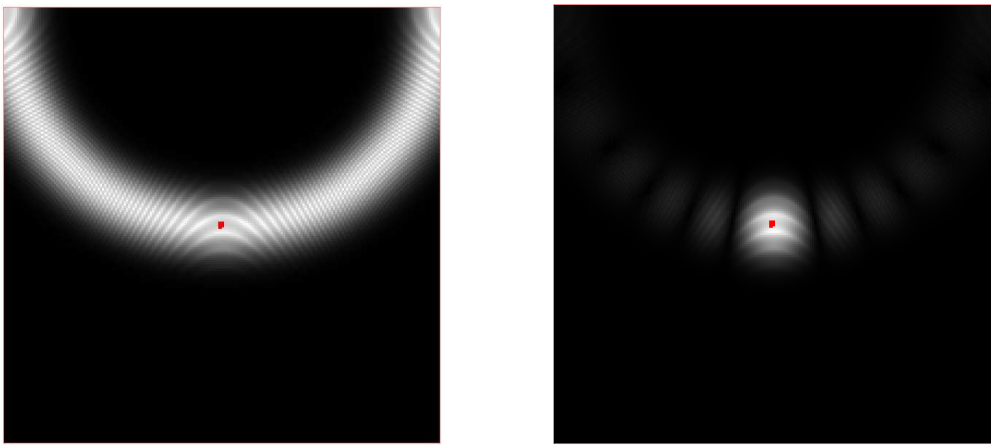


Figure 6.1: Simulation enviroment with two sources

To test the directivity of the microphone array, we use a narrow band pulse. We placed the pulse in the middle of the room and calculated the magnitude of signal for each point in the room. Points with higher magnitude are colored lighter as illustrated in Figure 6.2. In Figure 6.2a wave is omnidirectional because only one sensor exists. In Figure 6.2b the response of an array of 41 elements is depicted. Here we see a sinc pattern in θ domain. This phenomenon was discussed in equation 4.6.

Lobe containing the maximum power is called *main lobe* [27]. In figure 6.2b main lobe is lightest area with red point in the middle. The other lobes are called *side lobes*.



(A) Directivity of one microphone (B) Directivity of a microphone array

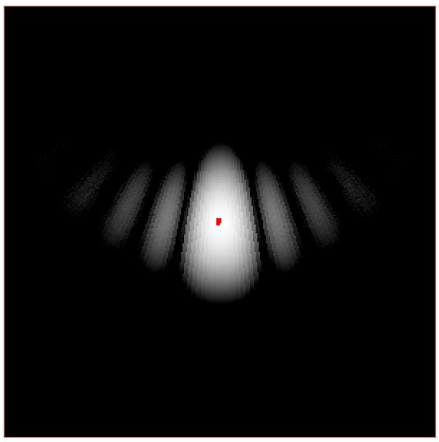
Figure 6.2: Directivity responses for a single sensor and 41 sensor microphone array

When element spacing is greater than a half wavelength, the spatial aliasing effect causes side lobes to become substantially larger in amplitude, and approaching the level of the main lobe; these are called *grating lobes*, and they are identical, or nearly identical copies of the main beams.

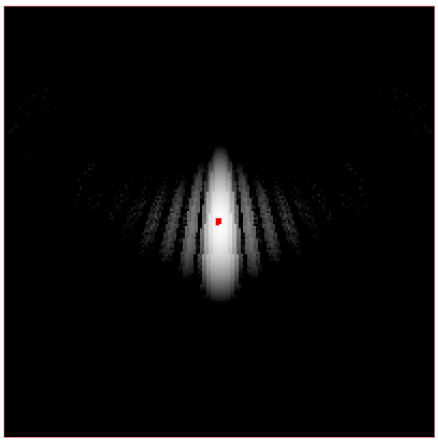
6.1.2 Effect Of Frequency On Directivity Pattern

Directivity response of microphone array to various frequencies is illustrated in Figure 6.3. There are 41 elements placed on microphone array and distance between elements are set to 10 cm. Red point in the simulation is spatial position of the pulse.

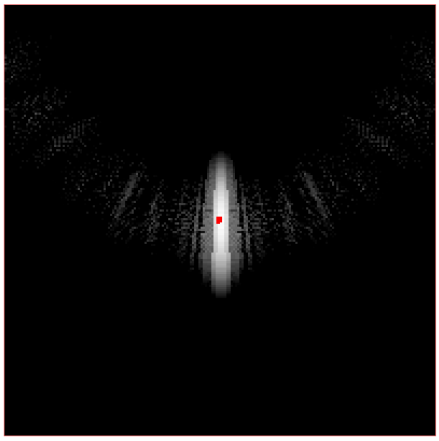
As can be seen frequency increase of the pulse increases the directivity gain of the array. On the other hand, we observe in Figure 6.3d that higher frequencies cause spatial aliasing. Three grating lobes can be identified in Figure 6.3d. The mathematical description of such behavior can be derived from Equation 2.7. Given distance between elements is 10 cm, maximum frequency of pulse that we can use without spatial aliasing effect is manually found as 1716Hz.



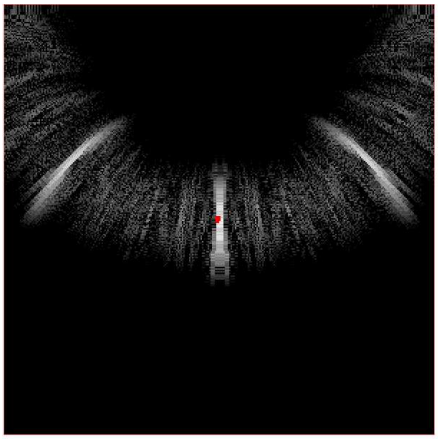
(A) Frequency set to 500Hz



(B) Frequency set to 1000Hz



(C) Frequency set to 1500Hz



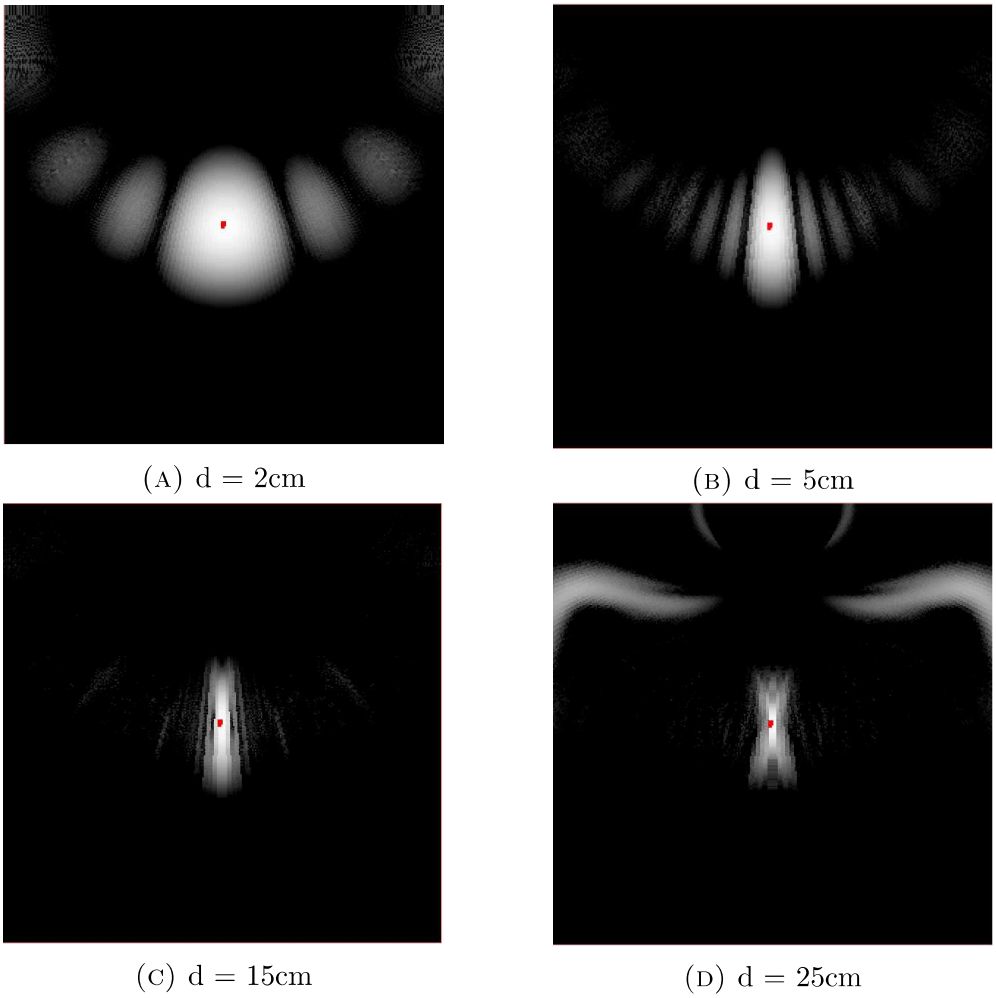
(D) Frequency set to 5000Hz, spatial aliasing occurred

Figure 6.3: Directivity responses different frequencies

6.1.3 Effect Of Distance Between Microphones

It can be deduced from Equation 4.6 that the size of the microphone array will increase the directivity response. We can add new elements to array and increase its size. This will be the subject of the coming section. We can also increase the size of the array with increasing the distance between microphones.

To show the effect of distance between microphones we ran simulations with different element distance values. It can be seen in Figure 6.4 that increasing the aperture distance will increase directivity until spatial aliasing occurs (Figure 6.4d).



(A) d = 2cm (B) d = 5cm

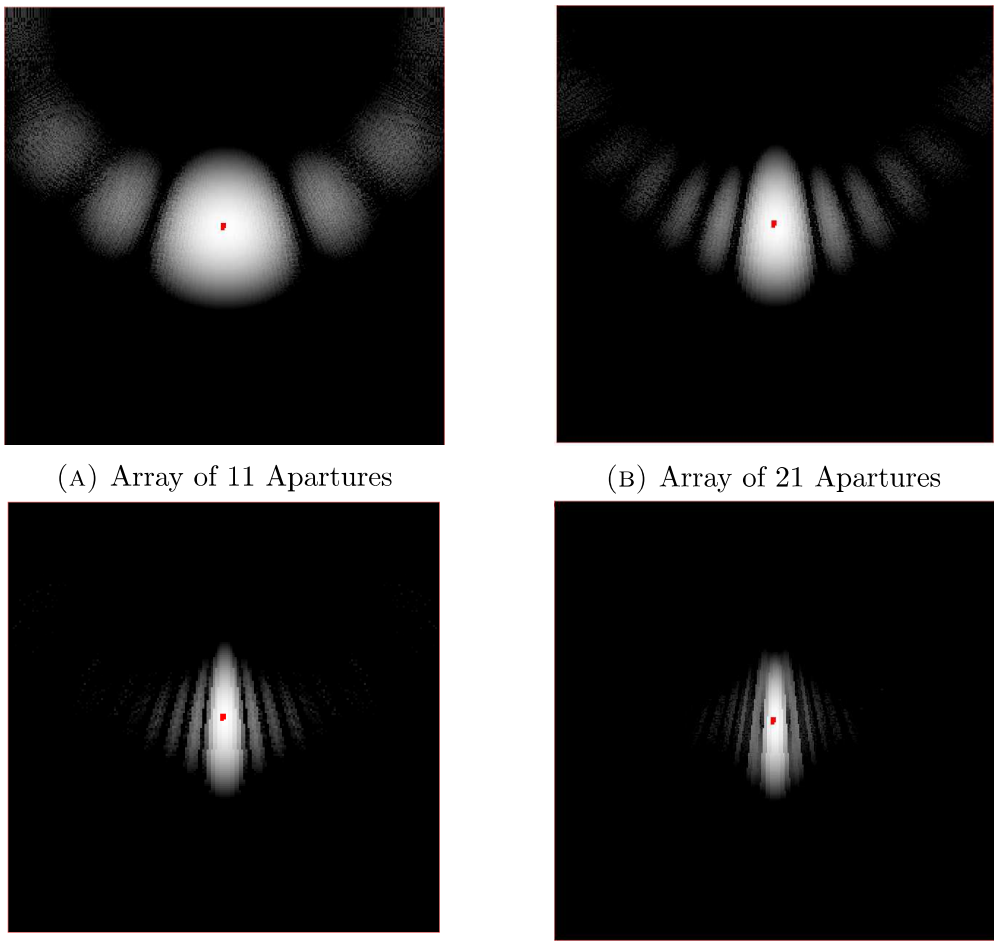
(C) d = 15cm (D) d = 25cm

Figure 6.4: Directivity responses to various distances between microphones

6.1.4 Effects of Aperture Count

Increasing window size of the filter kernel will increase the performance of the FIR filter. The same effect will be observed in microphone array element count. We saw that spatial aliasing imposes a limit on the frequency and distance between microphones. We don't have such limitation for element count. Increasing element count will always increase the microphone array performance. In this case the trade off will be the cost of the system and system complexity.

Directivity of several different microphone array configurations can be seen in Figure 6.5. It can be seen that the directivity gain increases with the element count.



(A) Array of 11 Apartures

(B) Array of 21 Apartures

(C) Array of 41 Apartures

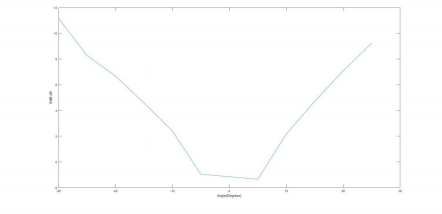
(D) Array of 61 Apartures

Figure 6.5: Directivity responses with different aparture counts

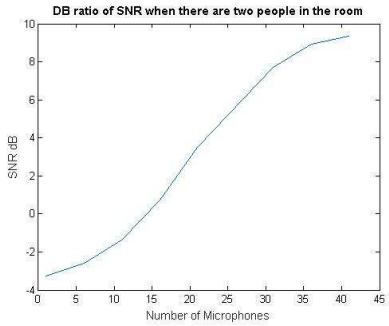
6.2 Microphone Array Experiments and Results

SNR is calculated and illustrated in Figure 6.6 with different configurations. While performing these tests target speaker is placed perpendicular to aperture and 10 meter away. An array with 41 elements used and distance between elements is selected as 10cm. In first configuration illustrated in Figure 6.6a interference speaker is assigned 10 meter away as well but with varying direction of arrival (DOA). It can be observed when interference is closer to the target speaker spatially, SNR values are dropping. Aperture performances are also compared with varying number of elements in 6.6b. While testing varying number of microphone array elements, target speaker is placed perpendicular to aperture

and 10 meter far-off. Interference speaker is placed on 0.27π direction and 10 meter far-off as well.



(A) SNR results while interference speaker is moving in DOA



(B) SNR results while increasing the reciever count

Figure 6.6: Delay-sum beamformer gain perfomances

6.3 Localization Experiments and Results

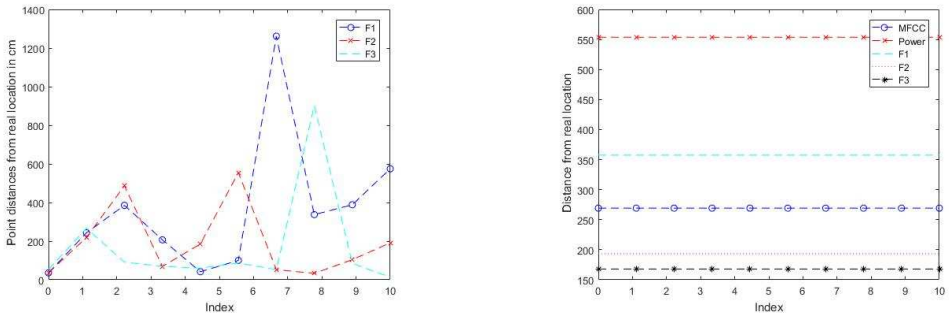
Until this point microphone array and speaker identification modules were considered as standalone applications. In this section both speaker identification and microphone array will be combined to improve the localization of the identified speaker.

Target speaker is assigned to a random location in our simulation room and an interference speaker assigned a random location within six meters radius of speaker's location. By using speaker identification and microphone array processing speaker's location is estimated. Finally euclidean distance in cm scale between real location and estimated location calculated. This process is repeated and results are illustrated in Figure 6.7 .

While estimating the location of the speaker several methods are used. Firstly we used signal power to estimate the location , than classical MFCC-GMM method and finally our offered method N-gram Formant methods are used. While estimating the location with signal power spatial filtering applied each point in the simulation environment and the point which has the maximum signal power is selected. For statistical methods each point is focused with spatial filtering again

but this time a score is obtained. An extra step applied in N-gram models. The probability of a point is multiplied with the magnitude of the speaker's formant frequency in that point to take advantage of spatial filtering more effectively. Because performance of the microphone array increases with increasing frequency formants between $1000 - 2000$, $2000 - 3000$ and $3000 - 4000$ are experimented separately. In Figure 6.7b the mean distances of estimated locations from real locations of the speaker is shown.

While obtaining these results a microphone array with 51 elements is used and distance between array elements configured as 10cm. Since not only DOA but also the radius of target speaker desired to be located; to increase near-field radius such large microphone array is chosen.



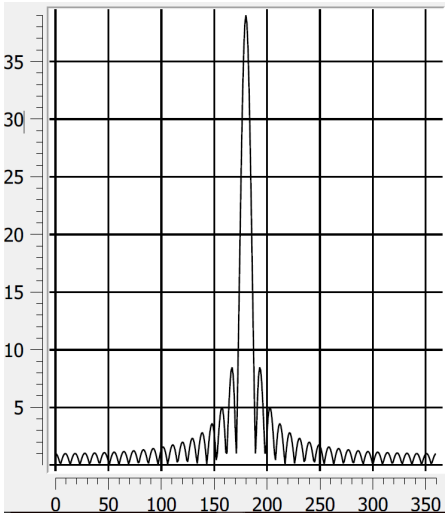
(A) Offered NGram Localization method's localization performance in cm

(B) Comparing various method's mean distances from real position in cm

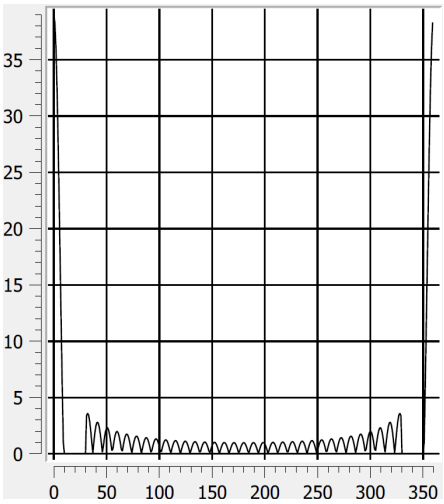
Figure 6.7: Localization experiments with varying identification methods

6.4 Pattern Shaping Results

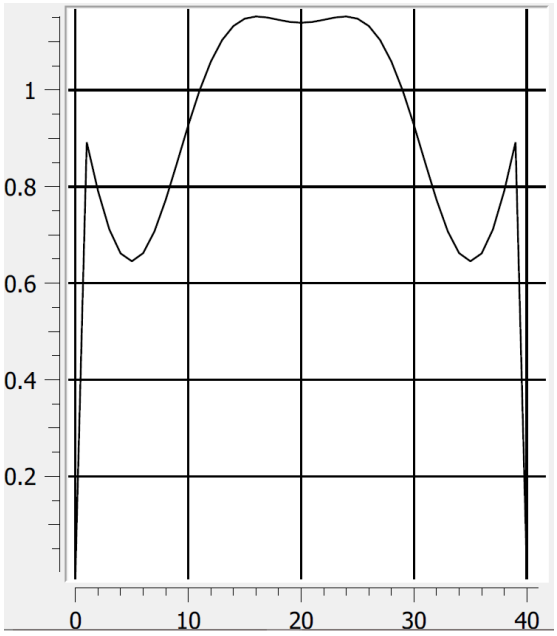
Process of pattern shaping(pattern nulling)[28] is illustrated in 6.8. Firstly as shown in 6.8a angular response of aperture is obtained. A window function created with nulls are assigned in interference DOA. Delay-sum microphone array's angular response is multiplied by window function. Result of second step is illustrated in Figure 6.8b. And finally IDFT operation applied to obtain aperture weight vector.



(A) Beam Pattern of a microphone array with 41 elements and d=10



(B) Windowing operation on interference direction



(C) Weight vector gathered from IDFT operation

Figure 6.8: Gathering weight vector from beam pattern

Weight vector which shown in 6.8 is applied when target speaker located perpendicular and interference's DOA is adjusted as 0.34π . SNR results are compared in Figure 6.9.

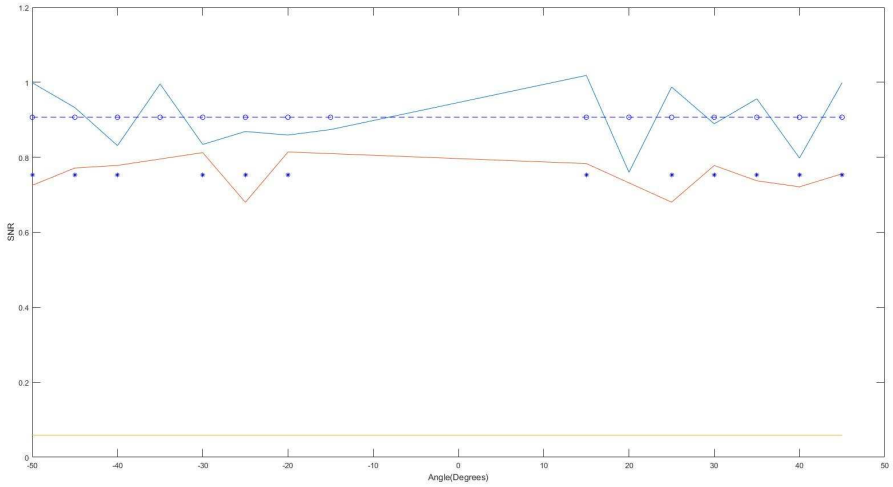


Figure 6.9: SNR result of pattern nulling beamformer and delay-sum beamformer. Pattern nulling beamformer is depicted as blue while delay-sum beamformer depicted as red. SNR values when there is no beamformer is depicted yellow in the buttom of the graphic.

Chapter 7

Conclusion

In this thesis we have sought to improve microphone array performance by using speaker identification methods. To apply advance microphone array methods, the desired angular response should be known, thus main problem of non adaptive microphone array methods is locating the target and interference locations. We have also focused to increase speaker identification performance with microphone array processing. Main problem of identification systems are decreasing accuracy with increasing noise. This makes reducing noise extremely important in identification systems.

Conventional microphone array processing algorithms calculates cross correlation between array elements to find the direction relative to the array. One location estimation problem is the cocktail party problem, where a number of people are talking simultaneously in a room. When only one specific person would like to be overheard in a cocktail party cross correlation methods will not be selective. Cocktail party problem is also a challenge for speaker identification systems. Blind source separation methods which uses only one receiver for speech separation can be used instead beamforming methods. While useful solutions can be derived, blind source separation problem is in general highly underdetermined.

The work presented in this thesis takes a different approach by considering the microphone array processor and the speech identification system as components of a single system. In this approach instead of looking cross correlation between

array receivers, speaker identification methods were used to estimate location of the target speaker. Iteratively to increase identification performance spatial filtering methods with microphone array were applied.

The contributions and findings of this thesis are summarized in the remainder of this chapter. Some remaining open questions about the work described in this thesis, and suggest some directions for further research are also explained.

7.0.1 Findings and Contributions

In our work we propose a identification method; $N - gramformant$, which uses features in high frequencies. The steps we have applied as follows:

- As conventional identification algorithms, Mel-frequency cepstral coefficients features were used in a Gaussian mixture model(GMM)
- To be able to calculate prosodic features fundamental frequency obtained by using Yins-FFT method.
- Since there is a formant roughly in each 1000Hz band, by using a simple approach; in each 1000Hz band, frequencies which are multiples of fundamental frequency were compared and the frequency which has the largest amplitude was selected.
- Formant vectors were used in a GMM to model target speaker's vocal tract shape.
- A scoring system were build with combination of speaker's probability and magnitude of the speaker's formant.

It was observed that using N-Gram and Formant models significantly increases the identification performance when used with MFCC feature. Also it was observed that given enough training data N-Gram model's accuracy increases up to

%80. Also the algorithm speeds of these models were compared with each other. Because of its simplicity N-Gram model were around 200 times faster than others.

At the same time we build a open source microphone array simulation environment [29]. In this simulation audio files can be assigned anywhere in the a $20m^2$ room. Firstly A delay-sum beamformer was build with in our simulation. After testing our identification system and microphone array simulation environment as standalone applications, they were combined to locate desired speaker with in a room. The steps we have applied as follows:

- An audio file which belongs to the target speaker was assigned to a location in the simulation environment.
- Noises were also inserted in the random locations.
- By using microphone array as a delay-sum beamformer, each location with in our simulation room was focused. Data gathered from each position is passed to identification system
- Data chunks were examined and scored by identification system where each chunk belongs to a spatial position. Chunk which target speaker located expected to give the best score.
- Chunk that has largest score was selected as speaker location.
- After speaker was located, interference was being searched around 10 meter radius of speaker's location. And location where had the largest energy level is selected as interference location

Location estimation by using classical MFCC-GMM and our offered identification method N-gram were compared. When N-gram method was used increase in location estimation accuracy was observed. Especially when the formant is selected in higher frequencies location estimation accuracy increased more. After speaker and interference locations were obtained, location information were used in IDFT

weight vector determination algorithm to increase microphone array performance. The steps we have applied as follows:

- A rectangular window function created. And nulls were assigned in interference DOA.
- Delay-Sum microphone array's angular response was multiplied by window function.
- IDFT applied on windowed angular response to obtain microphone array weight vector.
- Weighting applied in each element of microphone array.

Non-uniformly weighted and uniformly weighted beamformers performances were compared and it was observed that IDFT weight vector determination algorithm increases SNR.

7.0.2 Some Remaining Questions And Directions For Further Research

IDFT weight vector determination algorithm was applied while assigning microphone array element weights but methods which are often used like MVDR and LCMV beamformer have not been tried. Simulation environment can be extended that users can select which beamforming algorithm they want to use.

Beamforming was performed in time domain. This has been sufficient for narrow-band beamforming. But for implementing a broad-band beamformer a weight vector should be obtained for each frequency and applied in frequency domain.

Even threads were used some performance problems observed since identification is applied each data chunk from each spatial location. Using modern hardware like GPU can be a solution to performance problems.

Finally one of the biggest challenges in microphone array processing; $reflection$, is not simulated in our environment. For more realistic output $reflection$ should be added to simulation.

References

[1] Yiteng Huang Jacob Benesty, M. Mohan Sondhi. *Springer Handbook of Speech Processing*, chapter 3, page 39. Springer, 2008.

[2] B. Widrow. A microphone array for hearing aids. *The Journal of the Acoustical Society of America, vol. 109, issue 5*, pages 2426–2426, 05/2001.

[3] Dr. John McDonough Dr, Matthias Woelfel. *Distant Speech Recognition*, chapter 13, pages 416–419. Wiley, 2009.

[4] Iain McCowan. Microphone arrays: A tutorial. *Queensland University, Australia*, pages 1–38, 2001.

[5] Jacek Dmochowski, Jacob Benesty, and Sofiène Affès. On spatial aliasing in microphone arrays. *Signal Processing, IEEE Transactions on*, 57(4):1383–1395, 2009.

[6] Sverre Holm and Kjell Kristoffersen. Analysis of worst-case phase quantization sidelobes in focused beamforming. *Ultrasonics, Ferroelectrics, and Frequency Control, IEEE Transactions on*, 39(5):593–599, 1992.

[7] Walter Kellermann. A self-steering digital microphone array. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pages 3581–3584. IEEE, 1991.

[8] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1):19–41, 2000.

[9] Tomi Kinnunen and Haizhou Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, 52(1): 12–40, 2010.

[10] Anil K Jain, Arun Ross, and Salil Prabhakar. An introduction to biometric recognition. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(1):4–20, 2004.

[11] Robert J Vogt, Brendan J Baker, and Sridha Sridharan. Modelling session variability in text independent speaker verification. 2005.

[12] Li Deng. Front-end, back-end, and hybrid techniques for noise-robust speech recognition. In *Robust Speech Recognition of Uncertain or Missing Data*, pages 67–99. Springer, 2011.

[13] Frédéric Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, and Douglas A Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Advances in Signal Processing*, 2004(4):1–22, 2004.

[14] Claude C Chibelushi, Farzin Deravi, and John SD Mason. A review of speech-based bimodal recognition. *Multimedia, IEEE Transactions on*, 4(1): 23–37, 2002.

[15] Jacqueline Vaissière. Language-independent prosodic features. In *Prosody: Models and measurements*, pages 53–66. Springer, 1983.

[16] Yiteng Huang Jacob Benesty, M. Mohan Sondhi. *Springer Handbook of Speech Processing*. Springer, 2008.

[17] Satoshi Imai. Cepstral analysis synthesis on the mel frequency scale. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83.*, volume 8, pages 93–96. IEEE, 1983.

[18] IR Titze. Principles of voice production prentice hall. *Englewood Cliffs, NJ*, 1994.

[19] Guy J. Brown DeLiang Wang. *COMPUTATIONAL AUDITORY SCENE ANALYSIS*. IEEE, 2006.

[20] Bishnu S Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *the Journal of the Acoustical Society of America*, 55(6):1304–1312, 1974.

[21] Dr.R. S. Kawitkar Hemlata Eknath Kamale. Vector quantization approach for speaker recognition. *IJCTEE*, 2008.

[22] Douglas A. Reynolds. Automatic speaker recognition using gaussian mixture speaker models. *THE LINCOLN LABORATORY JOURNAL*, 1995.

[23] Mohamed Elhafiz Mustafa Sayed Jaafer Abdallah, Izzeldin Mohamed Osman. Text-independent speaker identification using hidden markov model. *World of Computer Science and Information Technology Journal*, 2012.

[24] Barry D Van Veen and Kevin M Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE assp magazine*, 5(2):4–24, 1988.

[25] Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[26] L.S. White and S King. The eustace speech corpus. *http://www.cstr.ed.ac.uk/projects/eustace*, 2003.

[27] Chen Jingdong Jacob Benesty. *Study and Design of Differential Microphone Arrays*. Springer, 2012.

[28] Quazi Mohmmad Alfred, Kousik Bishayee, Tapas Chakravarty, and Salil Kumar Sanyal. A dsp based study of pattern nulling and pattern shaping using transform domain window technique. *Progress In Electromagnetics Research C*, 2:31–45, 2008.

[29] Kadir Erdem Demir. Sharpear a microphone array simulation, 2013. URL https://github.com/kerdemdemir/sharpEar.