Accuracy Analysis of Sound Source Localization using Cross-correlation of Signals from a Pair of Microphones

Pradeep Kumar Yadav
Roll no. 213EC6274

Department of Electronics and Communication Engineering
National Institute of Technology, Rourkela
Rourkela, Odisha, India
June, 2015

Accuracy Analysis of Sound Source Localization using Cross-correlation of Signals from a Pair of Microphones

Thesis submitted in partial fulfillment of the requirements for the degree of

Master of Technology

in

Signal and Image Processing

by

Pradeep Kumar Yadav

Roll no. 213EC6274

under the guidance of

Prof. Lakshi Prasad Roy

Department of Electronics and Communication Engineering

National Institute of Technology, Rourkela

Rourkela, Odisha, India

June, 2015

dedicated to my parents...

National Institute of Technology
Rourkela

DECLARATION

I declare that

1. The work contained in the thesis is original and has been done by myself under the supervision of my supervisor.
2. The work has not been submitted to any other Institute for any degree or diploma.
3. I have followed the guidelines provided by the Institute in writing the thesis.
4. Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.
5. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Pradeep Kumar Yadav

National Institute of Technology
Rourkela

CERTIFICATE

This is to certify that the work in the thesis entitled **"Accuracy Analysis of Sound Source Localization using Cross-correlation of Signals from a Pair of Microphones"** submitted by ***Pradeep Kumar Yadav*** is a record of an original research work carried out by him under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of Master of Technology in Electronics and Communication Engineering (Signal and Image Processing), National Institute of Technology, Rourkela. To the best of my knowledge, this thesis has not been submitted for any degree or academic award elsewhere.

Prof. Lakshi Prasad Roy
Assistant Professor
Department of ECE
National Institute of Technology
Rourkela

Acknowledgment

This work is one of the most important achievements of my career. Completion of my project would not have been possible without the help of many people, who have constantly helped me with their full support for which I am highly thankful to them.

First of all, I would like to express my gratitude to my supervisor **Prof. Lakshi Prasad Roy**, who has been the guiding force behind this work. I want to thank him for giving me the opportunity to work under him. He is not only a good Professor with deep vision but also a very kind person. I consider it my good fortune to have got an opportunity to work with such a wonderful person.

I am also very obliged to **Prof. K.K. Mahapatra**, HOD, Department of Electronics and Communication Engineering for creating an environment of study and research. I am also thankful to Prof. A.K. Swain, Prof. A.K. Sahoo, Prof. S. Meher, Prof. S. Maiti, Prof. D.P. Acharya and Prof. S. Ari for helping me how to learn. They have been great sources of inspiration.

I would like to thank all faculty members and staff of the ECE Department for their sympathetic cooperation. I would also like to make a special mention of the selfless support and guidance I received from PhD Scholar Mr. Bibhuti Bhusan and Mr. Dheeren Kumar Mahapatra during my project work.

When I look back at my accomplishments in life, I can see a clear trace of my family's concerns and devotion everywhere. My dearest mother, whom I owe everything I have achieved and whatever I have become, my beloved father, who always believed in me and inspired me to dream big even at the toughest moments of my life, and brother, who were always my silent support during all the hardships of this endeavour and beyond.

Pradeep Kumar Yadav

Abstract

Sound source localization plays a significant role in many microphone array applications, ranging from speech enhancement to automatic video conferencing in a reverberant noisy environment. Localization of a sound source by using an algorithm which estimate the location from which the sound is coming from, aids to address; a long time hands-free communications challenge. The steered response power (SRP) using the phase transform (SRP-PHAT) method is one of the most popular modern localization algorithms. The SRP-based source localizers have been proved robust, however, the methods face high computation complexity, making it unsuitable for real time applications and also fail to locate the sound source in inimical noise and reverberation conditions, especially when the direct paths to the microphones are unavailable. This thesis works on a localization algorithm based on discrimination of cross-correlation functions (CCFs) [1]. The CCFs are determined using the method called the generalized cross-correlation phase transform (GCC-PHAT). Using cross-correlation functions, sound source location is estimated by one of the two classifiers: Naive-Bayes classifier and Euclidean distance classifier. Simulation results have demonstrated that the localization algorithms [1] provide a good accuracy in reverberant noisy environment.

Keywords: Sound source localization, Cross-correlation function, GCC-PHAT, Naive-Bayes classifier, Euclidean distance classifier.

Contents

List of Figures

List of Acronyms

Acronym	Description
ADC	Analog-to-Digital Converter
CCF	Cross Correlation Function
CPSD	Cross-Power Spectral Density
CT	Continuous Time
DAC	Digital-to-Analog Converter
DAS	Digital Audio System
DT	Discrete Time
FFT	Fast Fourier Transform
GCC	Generalized Cross Correlation
GCC-PHAT	Generalized Cross Correlation Phase transform
GM	Gaussian Model
PDF	Probability Density Function
RIR	Room Impulse Response
SSL	Sound Source localization

Chapter 1

Introduction

Sound Source Localization

History

Motivation

Problem Description

Thesis Organization

Chapter 1

Introduction

1.1 Sound Source Localization

Hearing, one of the primary sensors of a human body, is also one of the most important gateways to the mind. Therefore, the fascination with sound and hearing is almost as old as recorded history. Sound engineering as a whole has been a subject of interest to both scientific and music community. The particular aspect sound detection and localization has been of greater significance to a research community for the alternatives that it provides in contrast to burdened RF domain. Apparently, the medium of RF and sound are two orthogonal carrier mediums in nature with respect to propagation, frequency of operation and scope of application. The constraints of RF spectrum with respect to bandwidth regulation and RF pollution provide an impetus to explore sound like a viable alternative in as many applications as possible.

Sound source localization (SSL) is crucial to most microphone array applications such as speech enhancement [2], hands-free speech recognition [3], video-conferencing [4], and human-computer interface [5]. Many methods for SSL using microphone arrays have given in the literature [6]. For instance, distributed microphone network [7] and blind multiple-input multiple-output filtering [8] have been used to localize a sound source. Energy measurements

based method is given in [9]. Algorithms based on particle filtering are used to locate and track a wideband acoustic source [10,11]. In practice, the most popular methods are the time-difference-of-arrival (TDOA) estimation methods [12], such as adaptive eigenvalue decomposition (AED) [13] and generalized cross-correlation phase transform (GCC-PHAT) [14]. It has been shown that the method based on steered response power (SRP) is more robust than that based on time-difference-of arrival estimation. One of the most popular modern localization algorithms is the steered response power using the phase transform (SRP-PHAT) [16], also known as global coherence field (GCF) [15].

Localization methods can be divided into active and passive methods. Active methods send and receive energy whereas passive methods only receive energy. Active methods have the advantage of controlling the signal they emit which helps the reception process. Drawbacks of an active method include that the emitter position is revealed, more sophisticated transducers are required, and the energy consumption is higher compared to passive systems. As compared to active methods, Passive methods are more suitable for surveillance purposes since no energy is intentionally emitted. This thesis focuses on passive acoustic source localization methods.

1.2 History

Localization has been an important task in the history of mankind. In the beginning of modern navigation, one could determine his/her position at sea by measuring the angles from the horizon of celestial objects at a known time. The angles were determined via measurements, e.g., using a sextant. The celestial objects angle above the horizon at a particular time determines a line of position (LOP) on a map. The crossing of LOPs is the location. Modern navigation and localization utilize mainly electromagnetic signals. The applications of localization include radio detecting and ranging (RADAR) systems,

global positioning system (GPS) navigation, and light amplification by stimulated emission of radiation (LASER) -based localization technology. Other means of localization include utilization of sound waves in, e.g., underwater applications such as the sound navigation and ranging (SONAR).

1.3 Motivation

The main motivation of acoustic source localization is fuelled by the contribution it plays in a number of applications such as robot processing, automatic video camera shooting in video conferences, hands free mobile communications in the enclosed areas, and much more related applications. The Indian military, in particular, the army, has been engaged in border monitoring, a fact that has warranted untiring deployment of military resources of human capital and material. The primary aim of such deployment towards border monitoring is to detect enemy intrusion with an aim to counteract and pre-empt snowballing situations by checking intrusion at an early stage. Typically, across a border, gun firing is a typical phenomenon and many a times, the direction of these gun shots reveal vital information about adversaries. Currently, the security forces do not employ any sensor setup to detect events with acoustic emissions. In the event of a concrete and a fairly accurate acoustic detection system, the task of monitoring would significantly ease on requirement of men and material in that the manpower can be channelized in specific direction and area of interest rather than a general deployment spanning across the entire area. This would mitigate the stress and the logistic pressure as well. This need motivated the academic interest in this field to explore the feasibility of an end product based on the acoustic detection.

1.4 Problem Description

The process of sound source localization (SSL) is illustrated in fig 1.1. The SSL problem is divided into four stages: sound emission, propagation, measurement, and localization. The first three stages represent the physical phenomena and the measurement taking place before the localization algorithm solves the source location. These stages are briefly discussed in the following subsections. This thesis focuses on the last stage and discusses signal processing methods to locate the sound source. When discussing solutions to a problem, it is useful to classify the type of problem. The prediction of results from measurements requires:
1. A model of the system under investigations, and
2. A physical theory linking the parameter of the model to the parameter being measured.

A forward problem is to predict the measurement parameter from the model parameter, which is often straightforward. An inverse problem uses measurement parameter to infer model parameters. An example of the forward problem would state: output the received signal at the given microphone location by using the known source location and the source signal. Assuming a free-field scenario, this would be achieved by simply delaying the source signal by the amount of sound propagation delay between source and microphone position and attenuating the signal relative to the propagation path length. The inverse problem would state: solve the source location by using the measured microphone signals at known locations. The example inverse problem is much more difficult to answer than the forward problem. Hadamards definition of a well-posed problem is:
1. A solution exists
2. The solution is unique

3. The solution depends continuously on the data

A problem that violates one or more of these rules is termed ill-posed. During this thesis, it will become evident that sound source localization is an ill-posed inverse problem in most of the realistic scenarios.

1.5 Thesis Organization

This thesis consists of a total of six chapter organized as below :

Chapter 1 : This chapter gives a brief introduction to the sound source localization techniques in different environment and various applications of the sound source localization.

Chapter 2: This chapter will give some basics of audio and speech processing, which includes capturing and converting the audio signal, segmentation and windowing function etc.

Chapter 3: This chapter will give the overview of the sound source localization using Naive-Bayes and Euclidean distance classifier.

Chapter 4: This chapter will describe the implementation part of the SSL system.

Chapter 5: The localization accuracy of the SSL system and the dependency of it on various parameters are presented in this chapter.

Chapter 6: This chapter will discuss the conclusion and the future work.

Chapter 2

Audio and Speech Processing

Introduction

Capturing and Converting Sound

Noise

Signal

Sampling

Segmentation

Window Functions

Chapter 2

Audio and Speech Processing

2.1 Introduction

Audio and speech processing systems have steadily risen in importance in the everyday lives of most people in developed countries. From Hi-Fi music systems, through radio to portable music players, audio processing is firmly entrenched in providing entertainment to consumers. Digital audio techniques in particular have now achieved a domination in audio delivery, with CD players, Internet radio, MP3 players and iPods being the systems of choice in many cases. Even within television and film studios, and in mixing desks for live events, digital processing now predominates. Music and sound effects are even becoming more prominent within computer games. Speech processing has equally seen an upward worldwide trend, with the rise of cellular communications, particularly the European GSM (Global System for Mobile communications) standard. GSM is now virtually ubiquitous worldwide, and has seen tremendous adoption even in the worlds poorest regions.

2.2 Capturing and Converting Sound

Either sound created through the speech production mechanism or sound as heard by a machine or human, is a longitudinal wave which travels through air (or a transverse wave in some other media) due to the vibration of molecules. In air, sound is transmitted as a pressure variation, between high and low pressure, with the rate of pressure variation from low, to high, to low again, determining the frequency. The degree of pressure variation (namely the difference between the high and the low) determines the amplitude.

A microphone captures sound waves, often by sensing the deflection caused by the wave on a thin membrane, transforming it proportionally to either voltage or current. The resulting electrical signal is normally then converted to a sequence of coded digital data using an analogue-to-digital converter (ADC).

If this same sequence of coded data is fed through a compatible digital-to-analogue converter (DAC), through an amplifier to a loudspeaker, then a sound may be produced. In this case the voltage applied to the loudspeaker at every instant of time is proportional to the sample value from the computer being fed through the DAC. The voltage on the loudspeaker causes a cone to deflect in or out, and it is this cone which compresses (or rarifies) the air from instant to instant thus initiating a sound wave. In fact the process, shown

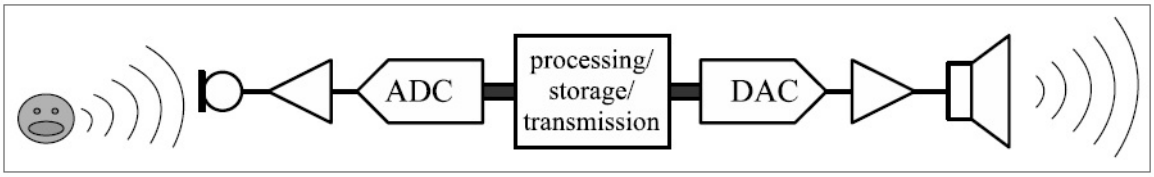


Figure 2.1: Block diagram of Digital Audio System

diagrammatically in figure 2.1, identifies the major steps in any digital audio processing system. Audio, in this case speech in free air, is converted to an

electrical signal by a microphone, amplified and probably filtered, before being converted into the digital domain by an ADC. Once in the digital domain, these signals can be processed, transmitted or stored in many ways, and indeed may be experimented upon using Matlab. A reverse process will then convert the signals back into sound.

2.3 Noise

Noise: Recorded sound is a combination of desired and undesired signals. The speech signal, we like to record is Desired signal; undesired signals include echoes, environment noise, and other speech signals.

2.3.1 Statistical Property

In the time domain, a noise signal has a good statistical model called a zero-mean Gaussian process $N(\mu_N, \sigma_N^2)$, with a known variance σ_N^2 and mean μ_N

$$p(y|\mu_N) = \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left(-\frac{(y-\mu_N)^2}{2\sigma_N^2}\right) \tag{2.1}$$

$$p(y) = \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left(-\frac{y^2}{2\sigma_N^2}\right) \tag{2.2}$$

The noise signals can be well described by the single variable (σ_N) Gaussian model (2.2). Figure (2.2) depicts the PDF of a noise, which is symmetrical about the mean $\mu = 0$ and has standard deviation $\sigma = 1$.

2.3.2 Spectral Property

Noise has a special kind of property which signifies each frequency contains how much power, is known as “Power spectral density” . The phase spectrum

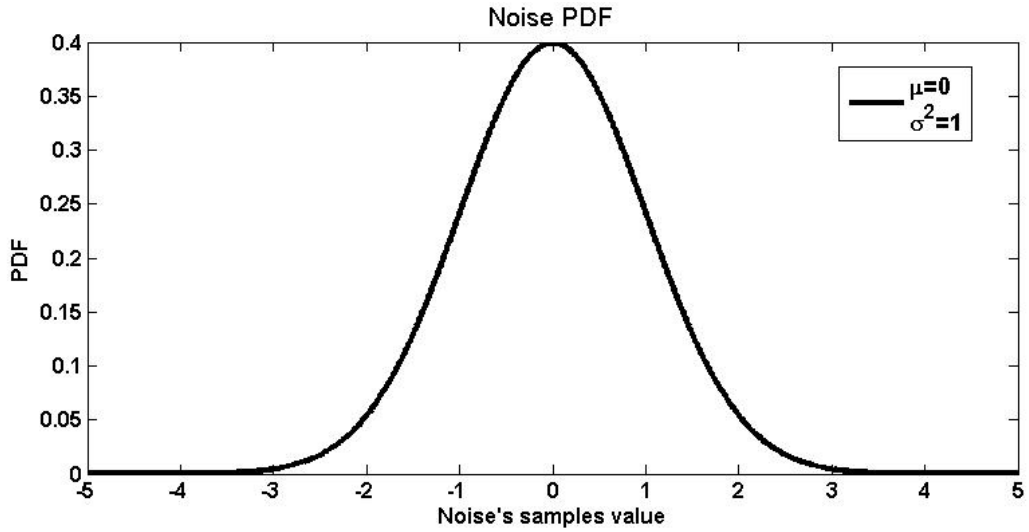


Figure 2.2: PDF of White Gaussian Noise

of white Gaussian noise is random in nature but the magnitude spectrum is flat i.e each frequency has an equal magnitude. Due to this reason, it is also called "white noise" – this is similar to the fact that the colour of the white light is white, due to the presence of all colours (frequencies).

The magnitudes of the practical noises in their spectrum usually decreases towards the high frequencies. A superior model for practical noises is the so-called "pink noise ", which has a slope of -6 dB/octave in magnitude spectrum at the high frequencies. White noise can be formed by passing the Pink noise through a low-pass filter of first order. The name Pink arrives again from an equivalence of light – red colour has a lower frequency and increasing the amount of lower frequencies leads to a more reddish colour.

"Blue noise" is one more kind of noise found in literature. The magnitudes of the blue noise increase 6 dB/octave towards the high frequencies. The equivalence with the light is same,i.e the blue noise has high magnitudes at high frequency as blue colour light has the highest frequency in visible spectrum. In the healthy environment, there is scarcity of the noises with spectral character-

istics same as blue noise.
A more kind of noise with the same analogy is called “coloured noise ”, which is a noise signal with a given magnitude spectrum (imitating a given colour) and random phases.

2.3.3 Temporal Property

One of the temporal properties of a noise signal assumes the noise is a stationary signal. This means that its probability distribution function and statistical properties in the time domain and mean and amplitude variation in the frequency domain, do not change with time. Sometimes, temporal property is referred to as “quasi-stationary ”, that is, the noise signal properties vary much more slowly than the other signal properties it is compared with.

2.3.4 Spatial Property

Spatial properties of a noise signal is given by the locations of the noise sources. In frequent cases, there is no firmly specified location of the noise source or there are too many noise sources.. In practice, the reverberation spreads the distinct location of the noise sources. Consequently, we are dealing with ambient noise in most of the cases; that means, there are no noise sources with firmly defined locations. In modelling, we can assume that the noise energy comes from every direction with equal probability, which is called “Isotropic ambient noise ”.

2.4 Signal

A signal is the desired part of the mixture of sounds recorded from the microphone. Unless stated clearly, by signal we will mean human speech. Human speech is a complex signal which comprises – a sequence of silence, plosive, unvoiced, and voiced segments.

• Silence segments are used to separate words and phonemes and are an integral part of human speech.
• In the signal spectrum, plosive segments contain transient variations of highly dynamic in nature .
•Unvoiced segments are noise-like turbulence, produced when the air is forced at high speed through a constriction in the vocal tract while the glottis is held open. They are characterized predominantly by the form of their spectral envelopes .
• Voiced segments are defined by their fundamental frequency, called "pitch " and its formants (harmonics). Basically, the pitch is the fundamental frequency of the vibration of vocal cords. The male voice has pitch frequencies range from 50 to 250 Hz, while female voice has pitch frequencies range from 120 to 500 Hz.

2.5 Sampling

2.5.1 Process

In nature, all signal present is a continuous time signal (CT-Signal) with continuous amplitudes. According to the theory, these signals have infinite time duration. In real systems, we cannot have all the amplitudes of a signal with infinite interval. Sound recording and processing system record the amplitude of the CT-Signal in particular moments of time. If these particular moments are of equal duration, then this equal duration is called "sampling period" T_s. Sampling period defines the sampling frequency of a CT-Signal as :

$$f_s = {}^{1}/_{T_s} \tag{2.3}$$

The process of changing a CT-Signal into a discrete time sequence (DT-Sequence) is called "Sampling". The result of this process is just a sequence of numbers.

The discrete time signal is represented as $x[n]$ where n is used as an index into the DT-Sequence. Mathematically, it can also be defined as :

$$x[n] = x(n\mathrm{T}_s) = x(t).\phi_{\mathrm{T}_s}(t) \tag{2.4}$$

where $\phi_{\mathrm{T}_s}(t)$ is the sampling function defined as :

$$\phi_{\mathrm{T}_s}(t) = \sum_{k=-\infty}^{+\infty} \delta(t - n\mathrm{T}_s) \tag{2.5}$$

where is called Dirac-delta functonand is defined as :

$$\delta(t) = \begin{vmatrix} 1 & t = 0 \\ 0 & \forall t \neq 0 \end{vmatrix} \tag{2.6}$$

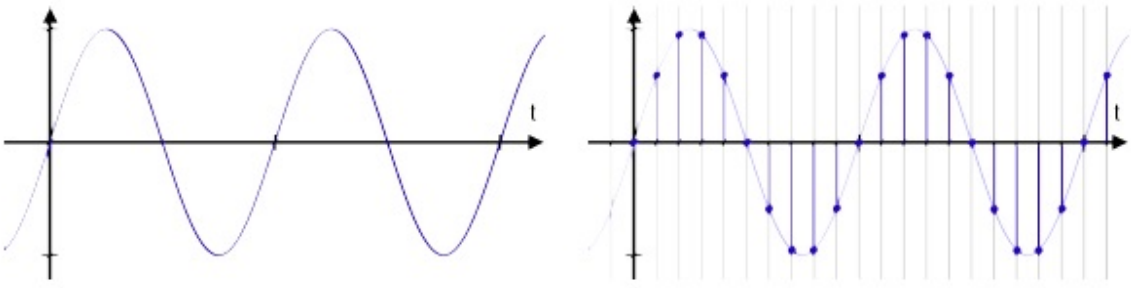


Figure 2.3: Sampling a sinosoidal signal

2.5.2 Theorem

The sampling theorem states the conditions under which a CT-Signal can be reconstructed accurately from its samples and moreover defines the interpolation algorithm that should be employed to get this accurate reconstruction. In simple terms, the sampling theorem states that the original continuous time signal can be reconstructed from its samples accurately, when the sampling frequency

f_s is greater than the twice of the maximum frequency f_m of the signal (seen as composition of sinosoids) to be discritized. :

$$f_s > 2f_m \tag{2.7}$$

The sample-rate is also called the Nyquist frequency, in honour to Harry Nyquist who also contributed a lot to sampling theory. When this condition arrives, we can reconstruct the underlying CT-Signal accurately by a process known as sinc-interpolation.

2.6 Segmentation

Segmentation is needed when any of the following are true:
1. The audio is continuous (i.e. we cant wait for a final sample to arrive before beginning processing).
2. The nature of the audio signal is continually changing, or short-term features are important (i.e. a tone of steadily increasing frequency may be observed by a smaller Fourier transform snapshot but would average out to white noise if the entire sweep is analysed at once).
3. The processing applied to each block scales nonlinearly in complexity (i.e. a block twice as big would be four or even eight times more difficult to process).
4. In an implementation, memory space is limited (very common).
5. It is desirable to spread processing over a longer time period, rather than performing it all at the end of a recording.
6. Latency is to be minimised a common requirement for voice communication systems.

Segmentation into frames is a basic necessity for much audio processing as mentioned above, but the process of segmentation does have its own fair share of problems. Consider an audio feature. By that I mean some type of sound that

is contained within a vector of samples. Now when that vector is analysed it might happen that the feature is split into two: half resides in one audio frame, and the other half in another frame. The complete feature does not resides in any analysis window, and may have effectively been hidden. In this way, features that are lucky enough to fall in the middle of a frame are emphasised at the expense of features which are chopped in half. When windowing is considered, this problem is exacerbated further since audio at the extreme ends of an analysis frame will be de-emphasised further. The solution to the lost-feature problem is to overlap frames.

2.7 Window Functions

Windows are the functions defined across the time record which are periodic in time record. They have starting and stoping value zero and are smooth function in between.

Speech is non-stationary signal where properties change quite rapidly over time. For most phonemes the properties of the speech remain invariant for a short period of time (5-100 ms). Thus for a short window of time, traditional signal processing methods can be applied relatively successfully. Most of speech processing in fact is done in this way: by taking short windows (overlapping possibly) and processing them. The short window of signal like this is called “frame”. In the windowing corresponds to what is understoods in filter design as window-method.

2.7.1 Hanning Window as an Example

Hanning window is the most widely used window. It has an amplitude variation of about 1.5 dB and provides better selectivity.

A window function is defined by a vector of real numbers $\{w_i\}$, $i = 1, 2, .., N-1$. It is used by multiplying the time series of any signal x_i with the window

before performing the FFT, i.e. using $x_i^{'} = x_i.w_i$ as input to the FFT.

$$w_i = \frac{1}{2}\left[1 - \cos\left(\frac{2\pi.i}{N-1}\right)\right];\ \ i = 1,2,...,N-1 \tag{2.8}$$

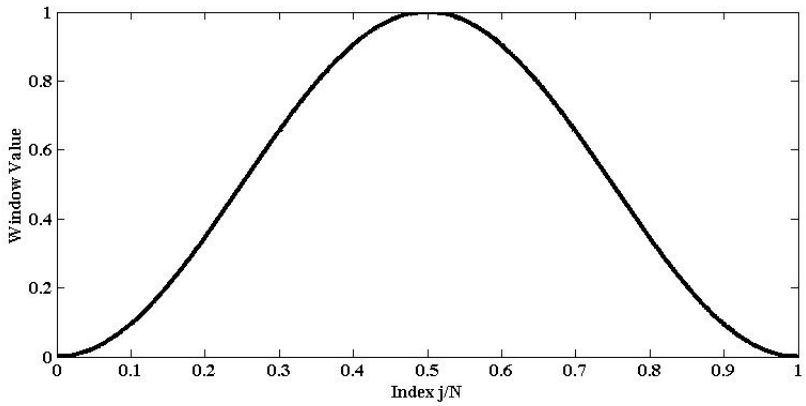


Figure 2.4: Hanning window

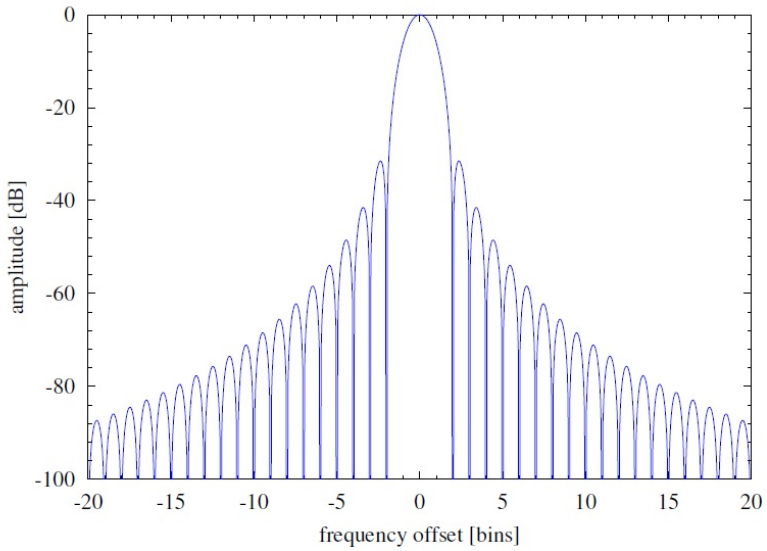


Figure 2.5: The Hanning window in frequency domain

Chapter 3

System Overview

Introduction

Block Diagram Of Sound Source Localization

Description of the Block Diagram

Chapter 3

System Overview

3.1 Introduction

The sound source localization system, going to be overviewed here, is a classification based localization system, in which location of sound source is estimated by one of the two classifier either Naive-Bayes classifier or Euclidean distance classifier.

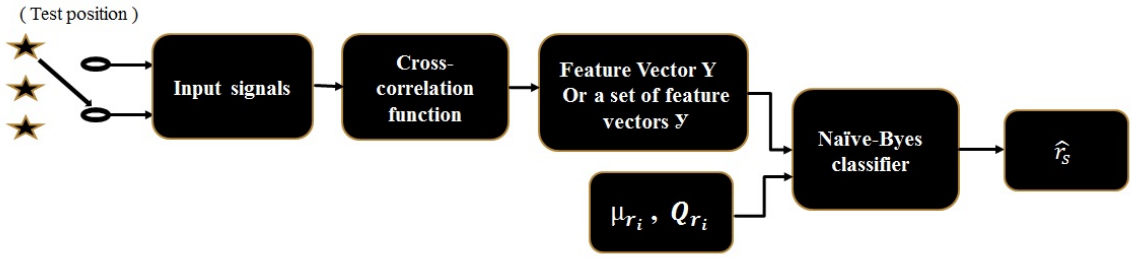
3.2 Block Diagram of Sound Source Localization

The block diagram of Sound Source Localization (SSL) consists of two phases:
1. **Training phase**
2. **Localization phase**

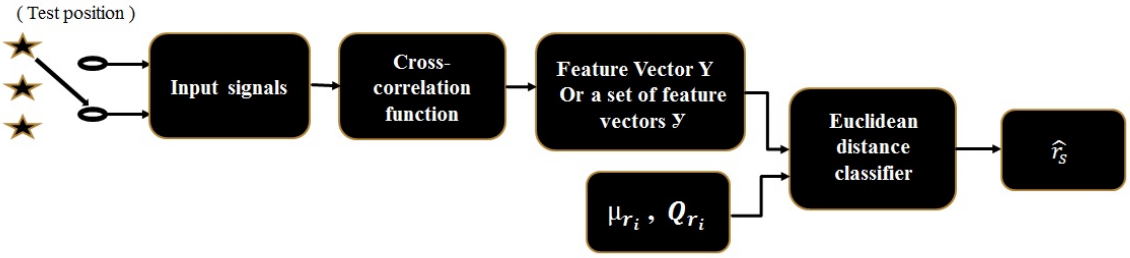
3.2.1 Training Phase of the Sound Sources

Figure 3.1: Training phase of each sound source

3.2.2 Localization Phase of a Sound Source



(a)



(b)

Figure 3.2: Localization phase of a sound source using (a) Naive-Bayes classifier (b) Euclidean distance classifier

3.3 Description of the Block Diagram

3.3.1 Training Phase

Fig.3.1 depicts the training phase of each possible source location (in fig. shown 'star') in a room. For this, we generate the two reverberant audio signals by convolving the RIRs of each source-microphone location pairs in the room. These signals are labelled as "Input signals" in the training phase of the block diagram of sound source localization. In order to get cross-correlation of these

two signals, we use generalized cross-correlation-phase transform method. Now, a feature vector Y is formed from the cross-correlated data. At the last, for each source location, mean vector and covariance matrix of cross–correlation functions are calculated and store to use in localization phase.

3.3.2 Localization Phase

Fig.3.2 depicts the localization phase of a sound source present in the room. When a reverberant signal (test data) comes from one of the trained locations, then the location of a source is estimated by discriminating the cross-correlation function. For this, the generalized cross-correlation function is calculated from the test data received at the microphones. Now, the feature vector (or a set of the feature vectors, when N frames of test data availabe) is formed by selecting the proper range of delay parameter of the generalized cross-correlation function. The last stage of the localization is source location estimation using one of the two classifier : Naive-Bayes classifier and Euclidean distance classifier .

3.3.2.1 Source Location Estimation using Naive-Bayes Classifier

When Naive-Bayes algorithm is employed, the classifier gives a good sound source location estimate based on the location l_k that maximizes the probability density function $p_{l_k}(\Upsilon)$.

Estimated sound source location $\hat{l_s}$ using Naive-Bayes classifier is given by:

$$\hat{l_s} = \arg\max_{l_k} p_{l_k}(\Upsilon) \tag{3.1}$$

where $p_{l_k}(\Upsilon)$ is determind by the mean vector μ_{l_k} and the covariance matrix C_{l_k}.

3.3.2.2 Source Location Estimation using Euclidean Distance Classifier

When Euclidean distance algorithm is applied, the classifier gives a good sound source location estimate based on the location that minimizes the Euclidean distance given in equation (4.24).

Estimated sound source location $\overset{\wedge}{l_s}$ using Euclidean distance classifier is given by:

$$\overset{\wedge}{l_s} = \arg\min_{l_k} \sum_{t=1}^{M} d_{l_k}^2 \left(\mathrm{y}^t\right) \tag{3.2}$$

where $d_{l_k}(\Upsilon^i)$ is Euclidean distance

$$d_{l_k}\left(\mathrm{y}^t\right) = \left(\left(\mathrm{y}^t - \mu_{l_k}\right)^T \left(\mathrm{y}^t - \mu_{l_k}\right)\right)^{1/2} \tag{3.3}$$

where "T" for transpose operation.

Chapter 4

Implementation

Reverberant Speech Signal Model

Cross-Correlation Function

Localization Algorithms

Chapter 4

Implementation

4.1 Reverberant Speech Signal Model

A reverberant speech signal is a combination of a speech signal and a reverberation. The reflection of the sounds from the walls of the room and other objects is called “reverberation”. A microphones signal is usually a reverberant signal, as it records not solely the direct speech signal from the source but also the multiple reflected signals that are the delayed versions of direct signal with changed spectrum and decreased magnitude .

The reverberation process is generally described by the room impulse re-

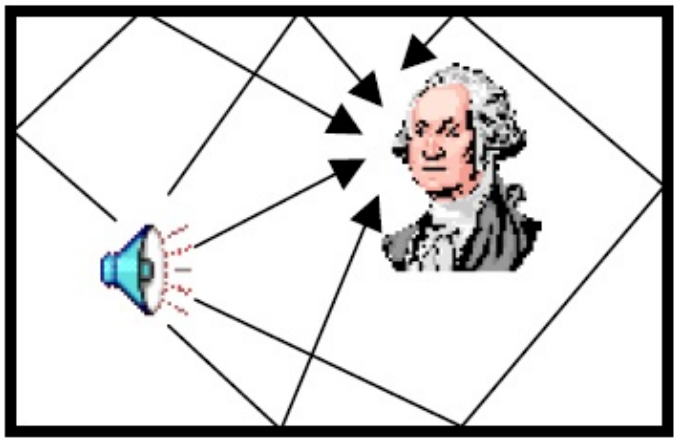
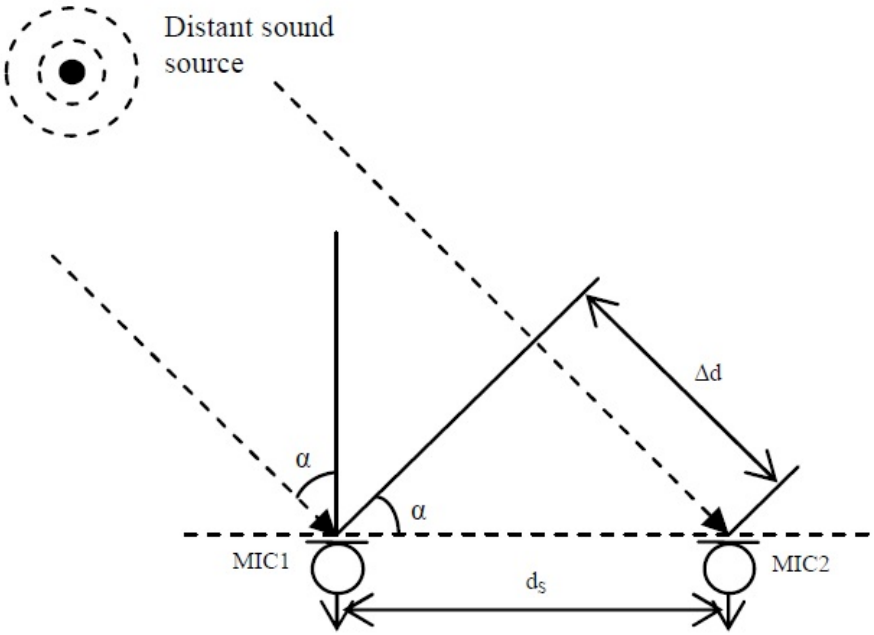


Figure 4.1: Reverberation in a room

sponse. In the array of two microphones as shown in figure 4.2, the reverberant speech signal received at the i^{th} microphone can be modelled as:

$$x_i(n) = h_i(l_s, n) * s(n), \qquad i = 1, 2 \tag{4.1}$$

where $s(n)$ is the discrete sound samples produced by the source located at l_s, $h_i(l_s, n)$ is the acoustic room impulse response from the source at l_s to the i^{th} microphone, and "*" denotes the convolution operation.



(b)

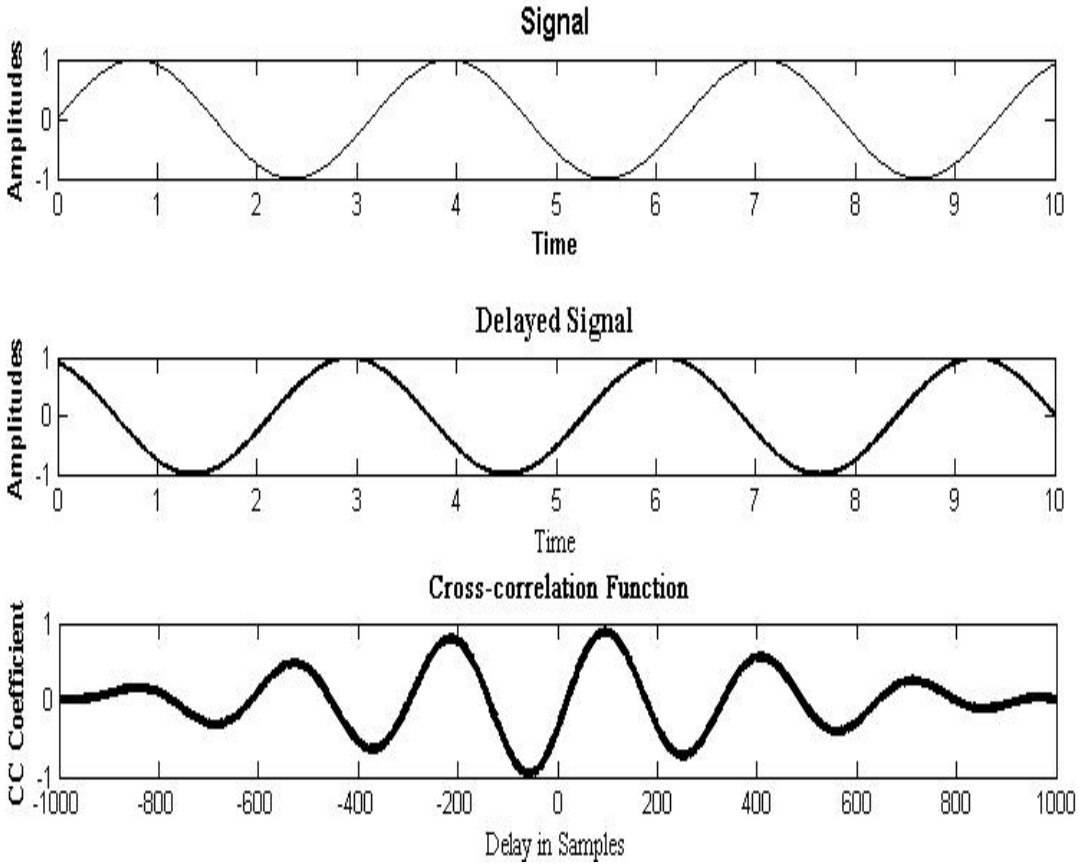
Figure 4.2: Sound wave reception

4.1.1 Acoustic Room Impulse Response

Acoustic room impulse response has all the information about the audible properties of the sound field. It is calculated using Image method of efficiently simulating small room acoustic [20],[21].

4.2 Cross-Correlation Function

The cross-correlation is a measure of similarity of two signals as a function of the delay of one with respect to the other. The peak value of cross-correlation function signifies the maximum similarity between the two signals at the certain dalay or lag, which is nothing but the delay between the two signals being compared.



(b)

Figure 4.3: Cross-correlation of the Sinosoidal signal with its delayed version

4.2.1 Determination of Cross-Correlation Function

The generalized CCF between the microphone-1′s signal $x_1(n)$ and the microphone-2′s signal $x_2(n)$ is determined in the freequency domain as :

$$R_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} \rho_{1,2}(\omega) X_1(\omega) X_2^*(\omega) e^{j\omega\tau} d\omega \tag{4.2}$$

where "*" denotes complex conjugation, $\rho_{1,2}(\omega)$ is the weighting function, and $X_1(\omega)$ and $X_2(\omega)$ are the Fourier transforms of the microphone signals $x_1(t)$ and $x_2(t)$ respectively.
In order to make generalized CCF less prone to reverberation, the phase transform method is used. By this method, the input signals are "whitened", which means the spectrum of input signals are uniformed with average magnitude.

$$\rho_{1,2}(\omega) = \frac{1}{|X_1(\omega) X_2^*(\omega)|} \tag{4.3}$$

Substituting (4.3) into (4.2), we have -

$$R_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} \frac{X_1(\omega) X_2^*(\omega)}{|X_1(\omega) X_2^*(\omega)|} e^{j\omega\tau} d\omega \tag{4.4}$$

4.2.2 Features of Cross-Correlation Function

In practice, the microphone signals $x_1(t)$ and $x_2(t)$ are firstly windowed using Hanning window (2.4), then the spectra $X_1(\omega)$ and $X_2(\omega)$ are calculated through Fourier transforms applied to these windowed segments. From (4.1), if the window length is greater than the length of the room impulse response, the spectra of microphone signals can be expressed as :

$$X_i(\omega) = H_i(l_s, \omega) S(\omega), \qquad i = 1,2 \tag{4.5}$$

where $S(\omega)$ and $H_i(l_s,\omega)$ are the Fourier transforms of $s(n)$ and $h_i(l_s,n)$ respectively.
Substituting (4.5) into (4.4), we have -

$$R_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} \frac{H_1(l_s,\omega)H_2^*(l_s,\omega)}{|H_1(l_s,\omega)H_2^*(l_s,\omega)|} e^{j\omega\tau} d\omega = R_{h_1h_2}(l_s,\tau) \tag{4.6}$$

From (4.6), the generalized CCF of the input signals $x_1(n)$ and $x_2(n)$ is equal to that of the room impulse responses $h_1(l_s,n)$ and $h_2(l_s,n)$. However, since the window length is shorter than the length of the acoustic room impulse response, the microphone signals' spectra are approximately expressed as :

$$X_i(\omega) \approx H_i(l_s,\omega)S(\omega) \tag{4.7}$$

and the generalized CCF of the input signals is approximately equal to that of the acoustic room impulse responses,

$$R_{x_1x_2}(\tau) \approx \int_{-\infty}^{\infty} \frac{H_1(l_s,\omega)H_2^*(l_s,\omega)}{|H_1(l_s,\omega)H_2^*(l_s,\omega)|} e^{j\omega\tau} d\omega = R_{h_1h_2}(l_s,\tau) \tag{4.8}$$

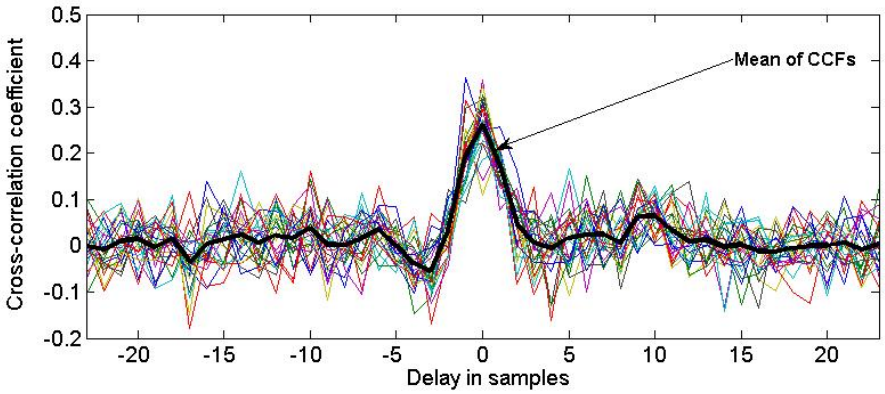


Figure 4.4: Cross-correlation Functions of the Source located at 90 degree from the mic-axis.

4.2.3 Modelling of Cross-Correlation Function

4.2.3.1 Feature Vector Formation

A Gaussian model is used to counteract the variability of the generalized-CCF. The generalized-CCF $R_{x_1x_2}(\tau)$ forms the feature vector

$$\mathrm{Y} \stackrel{\Delta}{=} \left[R_{x_1x_2}(-\tau_{\max}), R_{x_1x_2}(-\tau_{\max}+1), ..., R_{x_1x_2}(\tau_{\max}-1), R_{x_1x_2}(\tau_{\max})\right]^T$$

or

$$\mathrm{Y} \stackrel{\Delta}{=} \left[y_1, y_2, ..., y_j, ..., y_{2\tau_{\max}}, y_{2\tau_{\max}+1}\right]^T \tag{4.9}$$

where "T" denotes transpose operation, and

$$\tau_{\max} = round\left(\alpha D f_s / c\right) \tag{4.10}$$

where f_s is the sampling frequency, D is space between the microphones, c is the speed of sound, $round(\bullet)$ is rounding function, and the multiplication factor α is set to 1.67 in the next event.

4.2.3.2 Probability Density Function of Features

The probability density function of the features y_j, $j = 1, 2, ..., 2\tau_{\max}+1$ with the assumption that they are independently Gaussian with μ_j is the mean value, and σ_j^2 is the variance, can be given by equation (4.11):

$$p(y_j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(y_j-\mu_j)^2}{2\sigma_j^2}\right) \tag{4.11}$$

Fig. 4.5 shows the variations of the cross-correlation functions using simulated data.

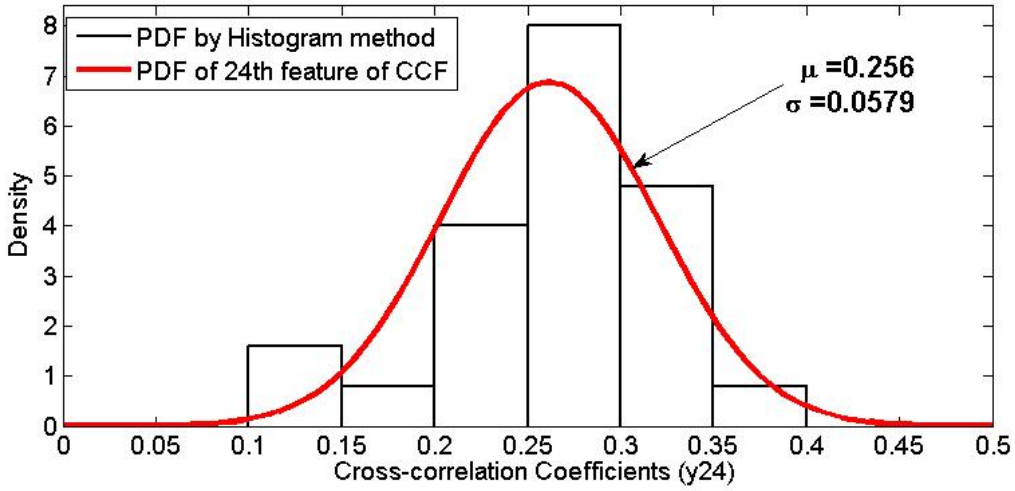


Figure 4.5: The estimated PDF of the feature y_{24}, and corresponding Gaussian PDF with the mean ($\mu_{24} = 0.256$) and standard deviation ($\sigma_{24} = 0.0579$).

4.3 Localization Algorithms [1]

4.3.1 Training of Features

The training of the feature is required in order to train the localization system. In this phase, the mean $\mu_j(l_k)$ and the variance $\sigma_j^2(l_k)$ of a feature for each location $l_k, k = 1,2,...,K$ are estimated using N frames training data.

$$\mu_j(l_k) = \frac{1}{N}\sum_{n=1}^{N} y_j^n \tag{4.12}$$

and

$$\sigma_j^2(l_k) = \frac{1}{N}\sum_{n=1}^{N} \left(y_j^n - \mu_j(l_k)\right)^2 \tag{4.13}$$

where y_j^n is the j^{th} feature of n^{th} training data.

4.3.2 Testing or Localization Phase using Classifier

4.3.2.1 Naive-Bayes Classifier

The Naive-Bayes classifier assumes that each feature is statistically independent from the other features in order to estimate the source location. It uses a feature vector y or a group of feature vectors $\Upsilon = \{\mathrm{y}^t, t = 1, ..., M\}$ of test data in order to estimate the location of the sound source using the mean vector μ_{l_k} and the covariance matrix C_{l_k} of the features from the training data.

Using these data of each location l_k i.e μ_{l_k} and C_{l_k} , we can calculate the PDF of each feature vector $p_{l_k}(\mathrm{y})$ of the test data as :

$$p_{l_k}(\mathrm{y}) = \prod_{j=1}^{2\tau_{\max}+1} p_{l_k}(y_j) = \prod_{j=1}^{2\tau_{\max}+1} \frac{1}{\sqrt{2\pi}\sigma_j(l_k)} \exp\left(-\frac{\left(y_j - \mu_j(l_k)\right)^2}{2\sigma_j^2(l_k)}\right)$$

or

$$p_{l_k}(\mathrm{y}) = \frac{1}{(2\pi)^{(2\tau_{\max}+1)/2}\left|C_{l_k}\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathrm{y} - \mu_{l_k}\right)^T C_{l_k}^{-1}\left(\mathrm{y} - \mu_{l_k}\right)\right) \quad (4.14)$$

where $\left|C_{l_k}\right|$ is the determinant of C_{l_k}. and

$$\mu_{l_k} = \begin{bmatrix} \mu_1(l_k) & \mu_2(l_k) & \cdots & \mu_{2\tau_{\max}+1}(l_k) \end{bmatrix} \quad (4.15)$$

$$C_{l_k} = \begin{bmatrix} \sigma_1^2(l_k) & & & \\ & \sigma_2^2(l_k) & & \\ & & \ddots & \\ & & & \sigma_{2\tau_{\max}+1}^2(l_k) \end{bmatrix} \quad (4.16)$$

In the case, when a group of the feature vectors $\Upsilon = \{\mathrm{y}^t, t = 1, ..., M\}$ is present,

then the PDF of Y can be given as :

$$p_{l_k}(\Upsilon) = \prod_{t=1}^{M} p_{l_k}(\mathrm{y}^t) = \prod_{t=1}^{M} \frac{1}{(2\pi)^{(2\tau_{\max}+1)/2}\left|C_{l_k}\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathrm{y}^t - \mu_{l_k}\right)^T C_{l_k}^{-1}\left(\mathrm{y}^t - \mu_{l_k}\right)\right) \tag{4.17}$$

$$p_{l_k}(\Upsilon) = \frac{1}{(2\pi)^{(2\tau_{\max}+1)M/2}\left|C_{l_k}\right|^{M/2}} \exp\left(-\frac{1}{2}\sum_{t=1}^{M}\left(\mathrm{y}^t - \mu_{l_k}\right)^T C_{l_k}^{-1}\left(\mathrm{y}^t - \mu_{l_k}\right)\right) \tag{4.18}$$

The last step is the estimation of sound source location $\overset{\wedge}{l_s}$ using Naive-Bayes classifier is given by:

$$\overset{\wedge}{l_s} = \arg\max_{l_k} p_{l_k}(\Upsilon) \tag{4.19}$$

4.3.2.2 Euclidean Distance Classifier

The Euclidean distance classifier is employed, when the variances of each feature are equal to σ^2 for each location $l_k, k = 1, 2, ..., K$

$$\sigma_j^2(l_k) = \sigma^2, \quad j = 1, 2, ..., 2\tau_{\max} + 1 \tag{4.20}$$

Substituting (4.20) into (4.16), the covariance matrix C_{l_k} can be rewritten as :

$$C_{l_k} = \begin{bmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \ddots & \\ & & & \sigma^2 \end{bmatrix}_{2\tau_{\max}+1 \times 2\tau_{\max}+1} \tag{4.21}$$

$$C_{l_k} = \sigma^2 \mathrm{I} \tag{4.22}$$

where "I" represents the Identity matrix of order $2\tau_{\max}+1$.
Substituting (4.22) into (4.18), we have -

$$p_{l_k}(\Upsilon) = \frac{1}{(2\pi)^{(2\tau_{\max}+1)M/2}\sigma^M} \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{M}\left(\mathrm{y}^t-\mu_{l_k}\right)^T\left(\mathrm{y}^t-\mu_{l_k}\right)\right)$$

or

$$p_{l_k}(Y) = \frac{1}{(2\pi)^{(2\tau_{\max}+1)M/2}\sigma^M} \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{M} d_{l_k}^2\left(\mathrm{y}^t\right)\right) \tag{4.23}$$

where $d_{l_k}(\mathrm{y}^t)$ is the Euclidean distance and is given by equation (4.24)-

$$d_{l_k}\left(\mathrm{y}^t\right) = \left(\left(\mathrm{y}^t-\mu_{l_k}\right)^T\left(\mathrm{y}^t-\mu_{l_k}\right)\right)^{1/2} \tag{4.24}$$

The last step is the estimation of sound source location $\overset{\wedge}{l_s}$ using Euclidean distance classifier is given by:

$$\overset{\wedge}{l_s} = \arg\min_{l_k}\sum_{t=1}^{M} d_{l_k}^2\left(\mathrm{y}^t\right) \tag{4.25}$$

Chapter 5

Simulation Results

Simulation Environment

Localization Accuracy

Dependency of Localization Accuracy

Accuracy Analysis

Chapter 5

Simulation Results

5.1 Simulation Environment

For simulation purpose, we have assumed a room of dimension $7 \times 6 \times 3$m as shown in figure 5.1. The walls of the simulation room have a frequency-independent and uniform reflection coefficient that does not change with the change in the direction or angle of the sound signal. The image method of

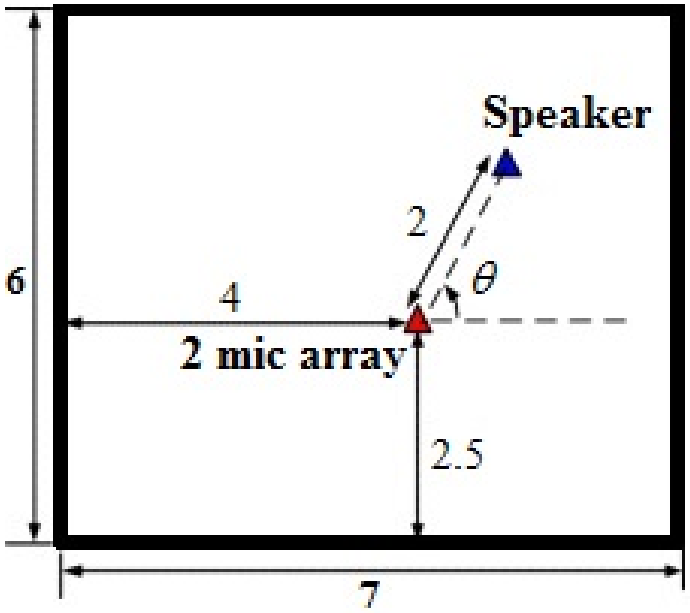


Figure 5.1: Simulation room

small room acoustic is used to generate the acoustic room impulse response. The positions of the two microphones are $x_1 = 3.85$, $y_1 = 2.5$, $z_1 = 1.2$ and $x_2 = 4.15$, $y_2 = 2.5$, $z_2 = 1.2$, where x, y, and z denotes the three different dimensions of the simulation room (all in meters).
The space between the two microphones is $D = 0.3m$ and the radial distance of the speakers from the mid-point of the line joining the two microphones is $2m$.

5.1.1 The Reverberant Speech Signal

The reverberant speech signals are made using "clean" speech and the acoustic room impulse response. When a "clean" speech signal is convolved with an acoustic room impulse response, then a reverberant signal is generated. The sampling frequency of the "clean" speech is 16 *khz*.

5.1.2 Adding White Gaussian Noise

The additive white noise has a characteristic that it is uncorrelated with the desired signal. Moreover, the additive Gaussian noises of the two microphones are uncorrelated with each other. In order to get different signal-to noise ratio (SNR), the reverberant signals are added with the zero mean white Gaussian noises of different variances. The SNRs of two microphones vary from 5 to 25 dB.

5.1.3 Segmentation and Windowing

The frame size of each reverberant noisy signal is 512 samples (32 ms), and each frame is windowed using a Hanning window given in equation (2.4).

5.1.4 Source Locations

We have considered seventeen speaker's locations 10^o, 20^o, 30^o, ..., 170^o for both training and testing.

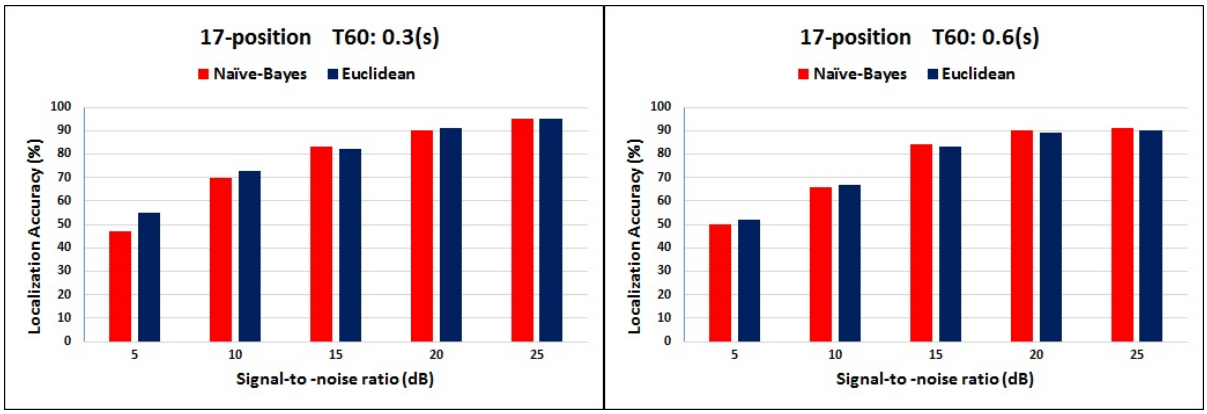
5.2 Localization Accuracy

An outcome of a localization system is labelled as true estimate when the localization outcome is the true speaker's location. The accuracy of a localization system is defined as the percentage of true estimates over all the localization estimates.

5.3 Dependency of Localization Accuracy

5.3.1 On the SNR and the Reverberation Time

Fig. 5.2 depicts the localization accuracy as a function of signal-to-noise ratio for the localization methods, where the number of training frames is 100 . As expected, at high signal-to-noise ratio levels, each of the methods performs very well. But as reverberation time increases, the performances for both algorithms become bad .

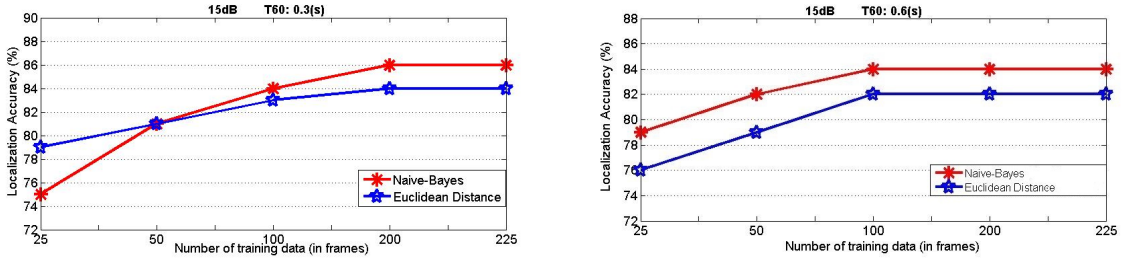


(a) (b)

Figure 5.2: Accuracy of localization system for the Naive-Bayes and the Euclidean distance algorithms with different SNR and Reverberation time (a) T60 = 0.3(s) (b) T60 = 0.6(s)

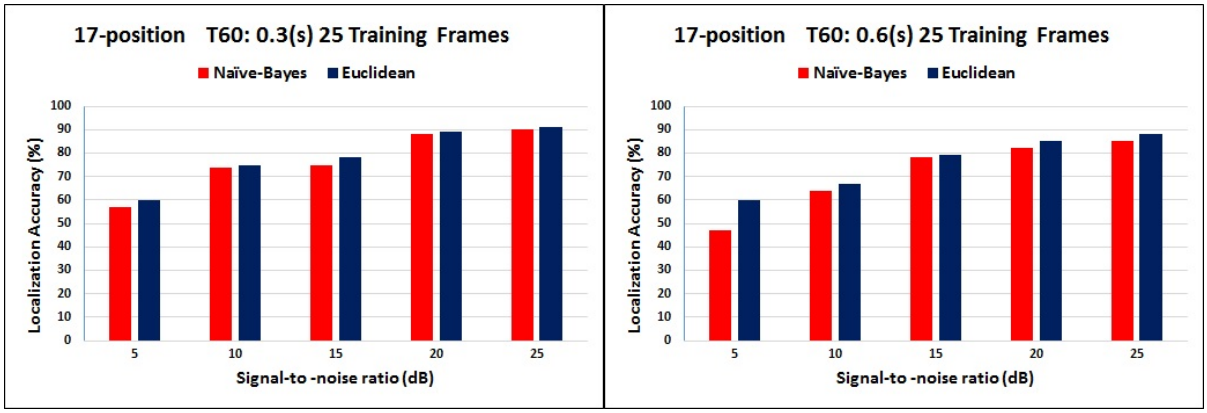
5.3.2 On the Number of Training Frames

Fig. 5.3 shows the dependency of the localization accuracy on the number of training frames for the two localization algorithms used. Fig. 5.4 depicts the localization accuracy where the number of training frames is 25, and Fig. 5.5 shows the localization accuracy where the number of training frames is 200. The performance of the training using 25 frames may be a bit poor due to the lack of data for estimating the mean values and covariance matrices of the generalized-CCF. The localization accuracy of the algorithms is improved by



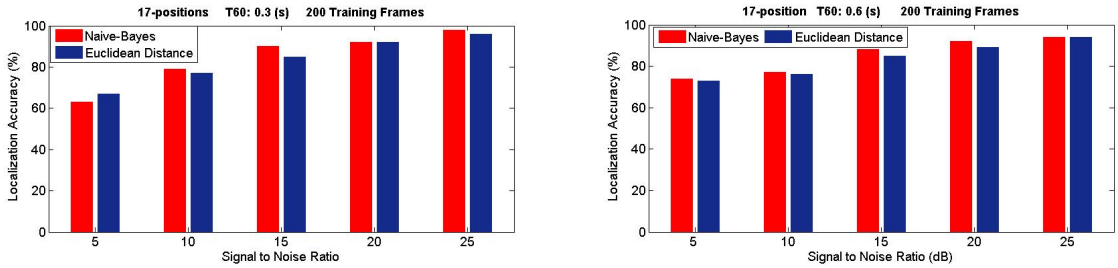
(a) (b)

Figure 5.3: Accuracy of localization system for the Naive-Bayes and the Euclidean distance algorithms with different number of training frames and Reverberation time (a) T60 = 0.3(s) (b) T60 = 0.6(s)



(a) (b)

Figure 5.4: Accuracy of localization system for the Naive-Bayes and the Euclidean distance algorithms with 25 training frames (a) T60 =0.3 (sec) (b) T60 =0.6 (sec)



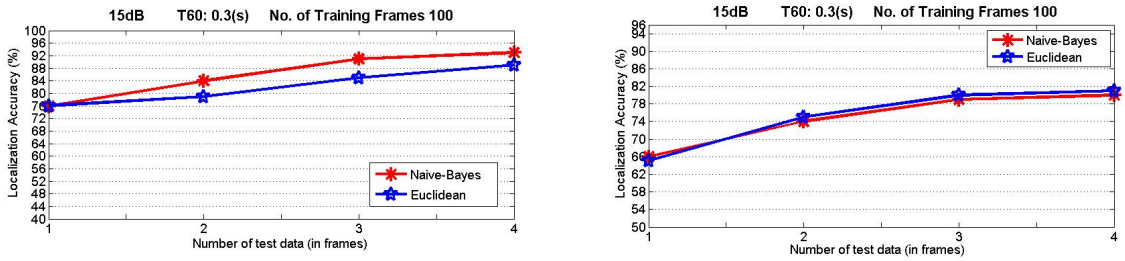
(a) (b)

Figure 5.5: Accuracy of localization system for the Naive-Bayes and the Euclidean distance algorithms with 200 training frames (a) T60 =0.3 (sec) (b) T60 =0.6 (sec)

increasing the number of training frames. When the number of training frames reaches 100, the localization accuracy does not increase significantly as the training frames' set size increases.

5.3.3 On the Number of Testing Frames

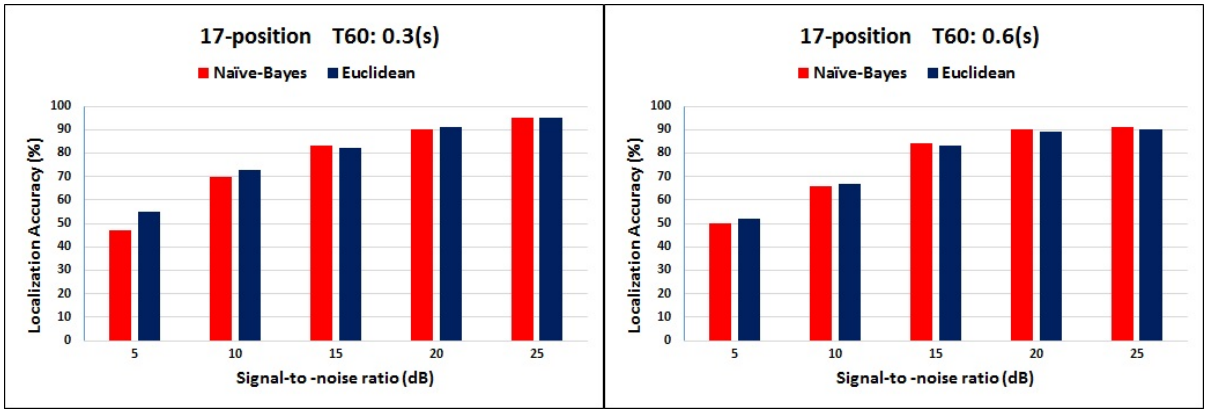
Fig.5.6 shows the localization accuracy as a function of the number of testing data for the localization algorithms [1], where the number of training frames is 100.



(a) (b)

Figure 5.6: Comparison of the performance for the Naive-Bayes and the Euclidean distance classifiers with various number of testing frames (a) T60 = 0.3(s) (b) T60 = 0.6(s)

Fig. 5.7 shows the localization accuracy where the number of testing data is two frames. It can be observed from Fig. 5.6 that the localization accuracy of the localization algorithms increases significantly as the testing set size increases.



(a) (b)

Figure 5.7: Accuracy of localization system for the Naive-Bayes and the Euclidean distance algorithms with two testing frames (a) T60 =0.3 (sec) (b) T60 =0.6 (sec)

5.4 Accuracy Analysis

5.4.1 PDF of the Untrained Locations

The PDF value of an untrained location's feature vector is calculated by using the mean vector and the covariance matrix of one of the trained or known location's feature vector.

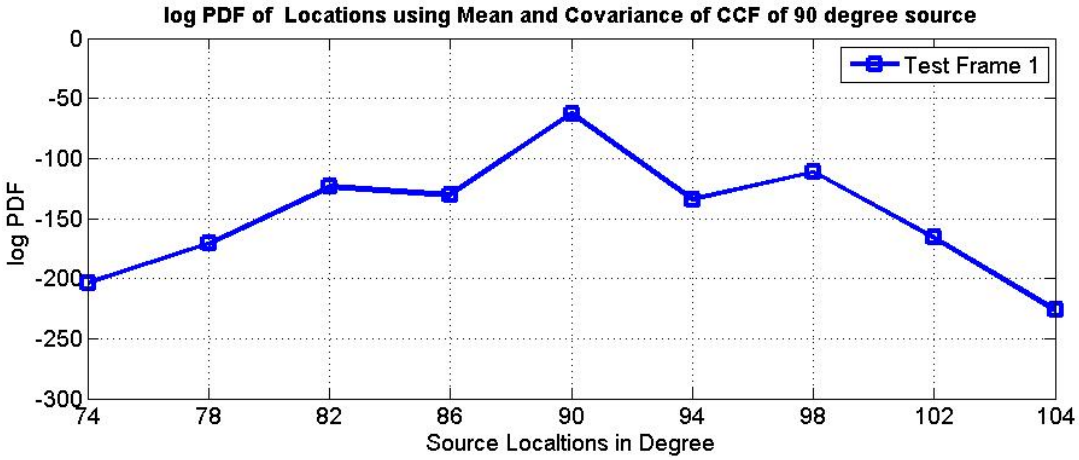


Figure 5.8: Untrained locations Vs log-PDF for number of test frame 1

By finding the PDF values for each untrained location, we have evaluated the numerical values of PDF through simulation and presented the results in fig. 5.8.

5.4.1.1 log-PDF Graph depends on the Number of Test Frame

Fig.5.9 shows the variation in log-PDF graph as a function of number of test frames. The difference between the log-PDF values of the trained location (90 degree source) and the neighbour location (92 degree) increases as the number of test frames increase. This increase in difference values of log-PDFs shows that the probability of true estimate increases with increase in the number of test frames.

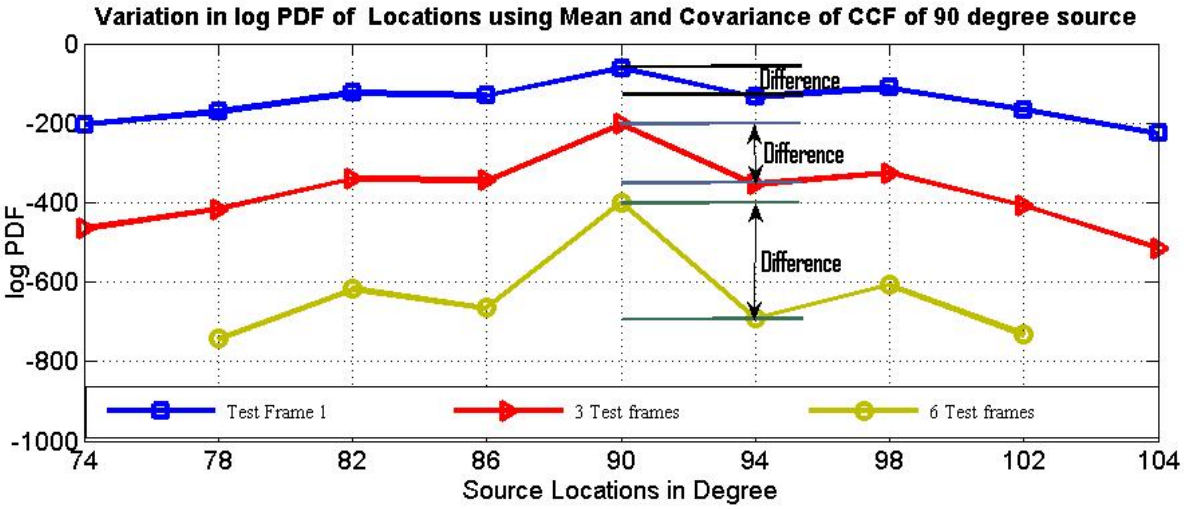


Figure 5.9: Untrained locations Vs log-PDF for different number of test frames

5.4.1.2 log-PDF Graph depends on the Reverberation Time (T60)

Fig. 5.10 shows the variation in log-PDF graph as a function of Reverberation time. The difference between the log-PDF values of the trained location (90 degree source) and the neighbour location (92 degree) increases as the reverberation time decreases. This increase in difference values of log-PDFs shows

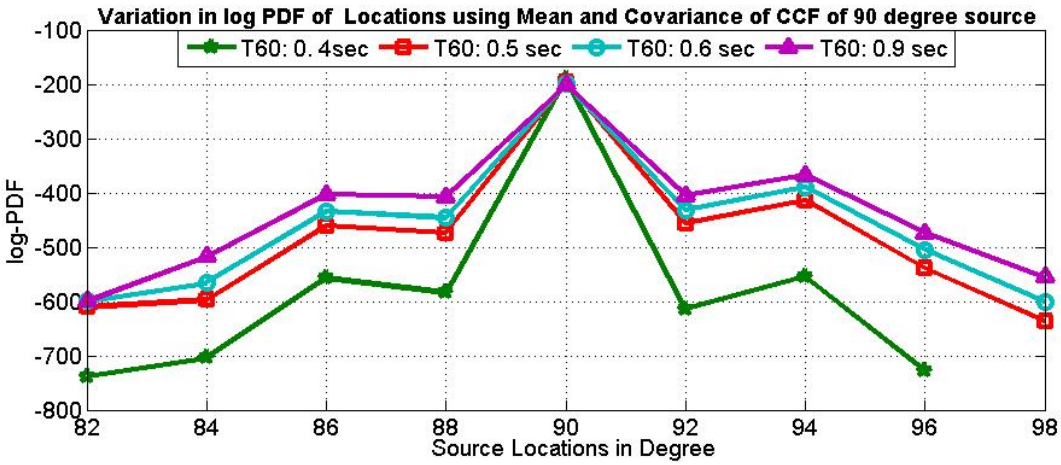


Figure 5.10: Untrained locations Vs log-PDF for different T60

that the probability of true estimate increases with decrease in the reverberation time.

5.4.2 Localization Accuracy of Naive-Bayes and Euclidean Distance Classifiers

Fig. 5.11 shows the comparison of the two algorithms using log-PDF graph.

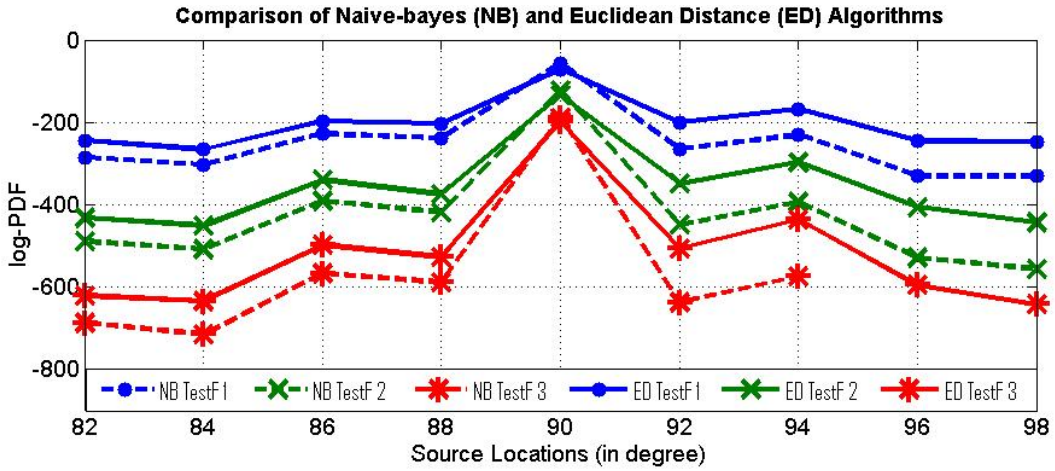


Figure 5.11: Comparison of Algorithms

It can be observed that the difference of log-PDF values of the trained and the untrained location is greater for Naive-Bayes algorithm than that for the Euclidean distance algorithm. Hence, Naive-Bayes method is a better discriminator than the Euclidean distance method. Moreover, as the number of test frames increases the performance of the Naive-Bayes algorithm becomes better than that of the Euclidean distance algorithm since the difference of log-PDF value becomes larger for Naive-Bayes algorithm than for Euclidean distance algorithm.

Chapter 6

Conclusion and Future work

Conclusion

Future work

Chapter 6

Conclusion and Future Work

6.1 Conclusion

We have simulated the SSL systems for both the algorithms [1] in MATLAB platform. These systems need the training of each location in a room, which is a very cumbersome process. In reverberant noisy environment, the localization accuracy of these systems is better than previous SRP-PHAT method based system as given in [1]. To overcome training process of each location, we have trained a particular location and in testing phase, we have calculated the PDFs of the feature vector using that trained location's mean vector and covariance matrix and have plotted these PDFs for different locations. This graph may be useful for the localization of untrained locations if we able to determine the relation between an untrained location's PDF value and the trained location's PDF value. The accuracy analysis for the two algorithms is presented in simulation results.

6.2 Future work

In our future work, we will try to locate the untrained locations by deriving the relation between the PDF of untrained locations and that of trained location.

Bibliography

[1] X. Wan and Z. Wu, "Sound source localization based on discrimination of cross-correlation functions," *Applied Acoustics*, vol. 74, no. 1, pp. 28–37, 2013.

[2] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2005.

[3] L. G. Brayda, C. Wellekens, M. Matassoni, and M. Omologo, "Speech recognition in reverberant environments using remote microphones.," in *ISM*, pp. 584–591, 2006.

[4] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2001.

[5] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey," *Proceedings of the IEEE*, vol. 98, no. 10, pp. 1692–1715, 2010.

[6] J. P. Dmochowski and J. Benesty, "Steered beamforming approaches for acoustic source localization," in *Speech Processing in Modern Communication*, pp. 307–337, Springer, 2010.

[7] G. Valenzise, G. Prandi, M. Tagliasacchi, and A. Sarti, "Resource constrained efficient acoustic source localization and tracking using a distributed network of microphones," in *Acoustics, Speech and Signal*

Processing, 2008. ICASSP 2008. IEEE International Conference on, pp. 2581–2584, IEEE, 2008.

[8] A. Lombard, H. Buchner, and W. Kellermann, "Multidimensional localization of multiple sound sources using blind adaptive mimo system identification," in *Multisensor Fusion and Integration for Intelligent Systems, 2006 IEEE International Conference on*, pp. 7–12, IEEE, 2006.

[9] K. Ho and M. Sun, "An accurate algebraic closed-form solution for energy-based source localization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2542–2550, 2007.

[10] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 826–836, 2003.

[11] Z. Liang, X. Ma, and X. Dai, "Robust tracking of moving sound source using scaled unscented particle filter," *Applied Acoustics*, vol. 69, no. 8, pp. 673–680, 2008.

[12] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP Journal on applied signal processing*, vol. 2006, pp. 1–19, 2006.

[13] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[14] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans Acoust Speech Signal Process*, vol. 24, no. 4, pp. 320–327, 1976.

[15] A. Brutti, M. Omologo, and P. Svaizer, "Comparison between different sound source localization techniques based on a real data collection," in *Hands-Free Speech Communication and Microphone Arrays, 2008. HSCMA 2008*, pp. 69–72, IEEE, 2008.

[16] J. DiBiase, H. Silverman, and M. Brandstein, "Microphone arrays. robust localization in reverberant rooms," *Microphone Arrays: Signal Processing Techniques and Applications*, 2001.

[17] X. Wan and Z. Wu, "Improved steered response power method for sound source localization based on principal eigenvector," *Applied Acoustics*, vol. 71, no. 12, pp. 1126–1131, 2010.

[18] A. Brutti, M. Omologo, P. Svaizer, and C. Zieger, "Classification of acoustic maps to determine speaker position and orientation from a distributed microphone network," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. IV–493, IEEE, 2007.

[19] N. Strobel and R. Rabenstein, "Classification of time delay estimates for robust speaker localization," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 6, pp. 3081–3084, IEEE, 1999.

[20] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J Acoust Soc Am*, vol. 65, no. 4, pp. 943–50, 1979.

[21] S. McGovern, *Room Impulse Response Generator*. http://in.mathworks.com/matlabcentral/fileexchange/5116-room-impulse-response-generator.