Implementation and Evaluation of a Low Complexity Microphone Array for Speaker Recognition

Prepared by: Peleira Nicholas Zulu

Supervised by: Prof. Daniel J. Mashao

University of Cape Town

Department of Electrical Engineering

December 2005

This dissertation is submitted to the University of Cape Town in fulfillment of the academic requirements for the Degree of Master of Science in Engineering.

Declaration

I declare that this thesis is my own work. It is being submitted for the degree of Master of Science in Engineering at the University of Cape Town. It has not been submitted for any degree or examination at any other university.

signature removed

Signature of Author: .....

Peleira Nicholas Zulu

December 2005

Acknowledgements

I would like to express my most sincere gratitude to my supervisor, Prof. Daniel J. Mashao for his guidance during the course of this research work. Without his invaluable advice, help and suggestions, this thesis would not have been possible. I am also particularly grateful for the many skills I acquired under his supervision. I would also like to thank him for the financial assistance I was offered during the course of my studies. Many thanks to my colleagues and friends in the STAR group, for all the laughs, support and useful suggestions. I would like to express my gratitude to my parents who have always been there for me, always encouraging me through good times and bad. Finally, I would like to thank God, for without Him none of this would have been possible.

Abstract

Hands-free operation is preferable in many potential speaker recognition applications. However, obtaining acceptable performance with a single distant microphone is challenging in real noise conditions. A possible solution to this problem would be to use microphone arrays.

Microphone arrays offer the possibility of hands free speech acquisition, which increases the convenience for users of speech technology applications as they do not need to hold a microphone in order to interact with the system. In addition, a microphone array has the advantage of potential gains in signal-to-noise ratio in noisy and reverberant environments. With the use of beamforming techniques, microphone arrays have the capacity to enhance a signal based purely on the knowledge of the direction of arrival of the signal.

This thesis discusses the application of a microphone array employing a noise canceling beamforming technique for improving the robustness of speaker recognition systems in a diffuse noise field. The microphone array enhanced speech is evaluated on distortion, signal-to-noise ratio and speaker recognition performance. The reported results show that the noise canceling beamformer with a post filter produces low distortion, high signal-to-noise ratio speech, and the best speaker identification and verification rates when compared with other general beamforming techniques.

Key words: Microphone array, beamforming, speaker recognition

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

ALC	- Adaptive Linear Combiner
DET	- Detection Error Trade-off
DS	- Delay and Sum
EER	- Equal Error Rate
FAR	- False Accept Rate
FRR	- False Reject Rate
GMM	- Gaussian Mixture Model
GSC	- Generalized Sidelobe Canceller
IS	- Itakura-Saito
LMS	- Least Mean Squares
LP	- Linear Prediction
MFCC	- Mel-frequency Cepstral Coefficient
MSE	- Mean Square Error
MVDR	- Minimum Variance Distortionless Response
NC	- Noise Canceling
PCI	- Peripheral Component Interconnect
PSD	- Power Spectral Density
SegSNR	- Segmental Signal to Noise Ratio
SID	- Speaker Identification
SNR	- Signal to Noise Ratio
SV	- Speaker Verification
TIMIT	- Texas Instruments and Massachusetts Institute of Technology
WNG	- White Noise Gain

Chapter 1

1 Introduction

Speech is the most natural form of communication for humans. For sometime it has been our goal to extend this form of communication to human-computer interaction to enable us to communicate with computers with the same ease and speed with which we communicate with other people. In addition, there are many situations where speech would be the most convenient way to communicate. Unfortunately, it is in those situations where speech technology would be most helpful that it performs the worst. For example, the high amounts of ambient noise present in an office environment would render laboratory speech and speaker recognition systems worthless.

Traditional speech and speaker recognition systems have been known to perform well when the speech signals are captured in a noise-free, single source environment using a close-talking microphone positioned near the mouth. However, many of the target applications of this technology do not take place in noise-free environments and it is often inconveniencing for the user to wear a close-talking microphone. As the distance between the speaker and the microphone increases, the speech signal becomes increasingly susceptible to background noise and reverberation effects that significantly degrade speech and speaker recognition accuracy [1]. As such, a need arises to design more robust systems that are capable of isolating or enhancing a desired signal from a mixture of spatially distributed signals. The primary focus of this thesis is to use a microphone array as a source localizing and signal enhancement speech acquisition system for speaker recognition.

1.1 Overview of microphone arrays

Microphone arrays provide a means of localizing sound pickup and improving sound quality in noisy and reverberant conditions [2]. A microphone array uses multiple spatially distributed sensors to capture speech signals. The speech signals are captured simultaneously by each of the microphones and then processed jointly using one or more of a variety of methods to obtain a cleaner output signal [3]. The most important objective of a microphone array is to provide a high quality version of the desired speech signal for a specified application. Microphone array speech enhancement techniques achieve this by what is referred to as *beamforming* [1-7], which reduces the level of localized and ambient noise signals, while minimizing distortion to speech from the desired direction. Beamforming is defined as the process of delaying and summing the outputs of multiple sensors in an array in order to reinforce a desired signal with respect to undesired signals propagating from different directions. The study and implementation of microphone arrays has been ongoing for years now and beamforming has been applied to speaker identification as in [7], using speech signals generated by a computer model of room acoustics. Beamforming algorithms utilize the spatial information of the noise and primary source signals to discriminate between the different signals. Increasing the number of microphones in an array and varying the beamforming algorithms used improves the ability of beamformers to extract the primary source using this spatial information. Previous work [8] has shown that microphone arrays can provide higher signal-to-noise ratios (SNR) than conventional microphones in distant talking environments.

1.2 Problem statement

Beamforming is one of the simplest and most robust means of spatial filtering, i.e., discriminating between signals based on the physical locations of the signal sources [9]. In a typical microphone array environment, the desired speech signal originates from a talker's mouth, and is corrupted by interfering signals such as other talkers and room reverberation. Spatial filtering can be useful in such an environment since the interfering signals originate from points in space other than the desired speaker's mouth [4]. Over the years researchers have developed various beamforming algorithms for microphone arrays for hands-free use in specialized applications. These

beamforming techniques can broadly be classified as being either data-independent or data-dependent [6]. *Data-independent* or *fixed* beamformers are so named because their parameters (such as the direction of the desired signal and the sensor weights) are fixed during operation. Conversely, *data-dependent* or *adaptive* beamformers continuously update their parameters based on the received signal. As different beamforming techniques are appropriate for different noise conditions [6], an initial investigation into the noise conditions encountered in the application target area is essential. This leads to the selection and design of an appropriate beamforming algorithm for the speaker recognition system.

Many studies aimed at evaluating different beamforming algorithms and microphone array geometries for various applications have been conducted, for example [1, 10, 11]. This thesis proposes and evaluates a beamforming technique that adaptively filters the incoming signals in order to pass the signal from the desired direction while attempting to reject noises coming from other directions. The technique integrates a specialized Generalized Sidelobe Canceller (GSC) beamformer and a Wiener post-filter for further enhancement. The beamformer is applied to improve speaker recognition performance in a distant talking office environment.

1.3 Thesis objectives

This thesis aims to resolve the problem of microphone array primary source enhancement in an office environment evaluated on speaker recognition. There are three main objectives encompassed in this thesis. The foremost objective is to provide a comprehensive review of existing microphone array literature with particular emphasis on how microphone arrays work, factors that affect their performance and beamforming algorithms that have previously been used to improve speaker recognition robustness.

The second objective of this thesis is to design and implement a microphone array system using ideas and techniques detailed in existing literature. The physical array structure will provide a means for signal capture while signal processing algorithms will be implemented on a computer. Single microphone (unprocessed) speech will provide a baseline to which the microphone array processed speech will be compared.

The baseline will act as the point of reference to which all beamforming algorithms and all improvements discussed in this thesis will be compared.

The final objective will be to implement the specialized generalized sidelobe canceller (noise canceling) beamformer with Wiener post-filtering. The performance of this beamformer will be compared to the baseline and other commonly used beamforming algorithms. The main evaluation of the system will be based for the most part on its performance on speaker recognition tasks.

It is not the aim of this thesis to present a complete solution to the problem of noise in speaker recognition systems, but simply to make a small contribution to the immense literature on the use of microphone arrays for speaker recognition that already exists.

1.4 Contribution to knowledge

One of the main objectives of this study is to improve speaker recognition robustness. This could be achieved in part by minimizing the problem of ambient noise in the speech signals. This is done by using a microphone array and a noise canceling beamformer with a post-filter. The noise canceller is usually associated with the generalized sidelobe canceller (GSC) beamformer and provides a set of filters that minimize the noise power in the output. The desired signal is initially blocked from a secondary path by a blocking matrix ensuring that only the noise power is minimized. The *least mean squares* (LMS) algorithm is a practical algorithm that permits us to find an approximated solution to the optimal filtering process. The resulting signal is then subtracted from the original signal. Thus the output of the beamformer contains the original signal with reduced noise power. In addition to this, a post-filter is added to further reduce the noise in the signal and to rectify any distortion caused by the beamformer. In so doing it is shown that speaker recognition performance is improved due to the reduction of both noise and distortion to input speech signals.

1.5 Scope and limitations

This chapter has so far presented a general overview of microphone arrays that encompasses a broad range of possible applications. However, the remainder of this

thesis is limited to exploiting the noise reduction potential of microphone arrays for the increased robustness of speaker recognition systems.

As mentioned previously, a technique employing a noise canceling beamformer and a post-filter is proposed to minimize noise and distortion in the input speech signals to a speaker recognition system. A variety of methods are suggested to overcome the effects of noise on speaker recognition systems. These can be broken into four general categories: feature classification, feature extraction, signal processing and signal acquisition. This thesis focuses on a technique applied at the signal acquisition level. Techniques at this level include the use of specialized microphones, multiple microphones with beamformers and other types of transducers, such as accelerometers.

The technique discussed in this thesis focuses on signal acquisition using multiple microphones with a beamformer. For this reason, only the beamforming technique used is compared with other commonly used techniques that also operate at the signal acquisition level. This is done to determine how the performance of the noise canceling beamformer with a post-filter compares with these techniques.

The evaluation of the noise canceling beamformer with post-filter is limited to speech degraded by ambient noise, that is, noise associated with the environment at a specified time, being usually a composite of sounds from many sources and directions. In particular, typical office noise which includes office machinery sounds like computer fans, disk drives, air conditioners, sounds generated by people moving around and background conversations. The robustness of the noise canceling beamformer with post-filter to degradations caused by other effects other than ambient noise was not evaluated.

Many actual systems using microphone arrays have been built and tested. In these systems, several parameters were considered. In addition to the complexity and computation issues, which are dependent on the algorithm(s) chosen and on the number and positioning of microphones in the array, some physical considerations must be made. A compromise must be made between a large microphone separation distance, which will provide good spatial resolution, and a small microphone separation distance, which better conforms to the *far-field* assumption. In addition,

this spacing between microphones must be less than half of the smallest wavelength of interest to avoid spatial aliasing.

A baseline system was developed to provide an experimental framework to which the noise canceling beamformer with post-filter could be compared for potential improvements.

1.6 Thesis Overview

Chapter 2 discusses background information of microphone arrays. Array geometries, characteristics and the algorithms used with microphone array beamformers are explained in detail. Chapter 3 presents a brief review of speaker recognition systems and the role of microphone arrays in this field. As previously mentioned, this research proposes a noise canceling beamformer modeled around the generalized sidelobe canceller. Chapter 4 introduces this technique, examining its mathematical formulation and noise field compatibility. To acquire the data used in this research requires the design of a microphone array data acquisition module. This is detailed in chapter 5 along with a brief description of the hardware and software used. Chapter 6 outlines the experimental results while chapter 7 discusses the results and gives direction for further research.

1.7 Summary

The aim of this chapter is to provide a brief introduction to the area of microphone arrays in order to familiarize the reader with the various terms and concepts used throughout the remainder of this document. Microphone array primary source enhancement for speaker recognition systems was highlighted as one of the main objectives to be addressed by this thesis. Other important aspects such as the objectives of this thesis, the scope and limitations and the thesis overview were also provided. The following chapter discusses the many fundamentals and concepts involved in the implementation of microphone array systems.

Chapter 2

2 Microphone Array Fundamentals

While the use of sensor arrays is relatively new in speech processing, the fundamental theory is well established as it is common to all sensor arrays, being based on the theory of wave propagation. Array processing involves the use of multiple sensors to receive or transmit a signal carried by propagating waves. Sensor arrays have applications in a variety of fields such as sonar, radar, seismology, radio astronomy and tomography [6][1]. This research focuses on the use of microphone arrays to receive speech signals. This section seeks to develop the principles of array processing by discussing some key features of discrete sensor arrays.

2.1 Wave propagation

Sound waves propagate through fluids[2] as longitudinal waves. The molecules in the fluid move back and forth in the direction of propagation, producing regions of compression and expansion. Using Newton's equations of motion to consider the infinitesimal volume of the fluid, an equation governing the wave's propagation can be developed. A generalized *wave equation* for acoustic waves is quite complex because it depends on the properties of the fluid. However, assuming an ideal fluid with zero viscosity, the wave equation can be derived as [12]

[1] Much of the information in this chapter was taken from this reference.
[2] Fluids include liquids, gases and plasmas

$$\nabla^2 x(t,\mathbf{r}) - \frac{1}{c^2}\frac{\delta^2}{\delta t^2} x(t,\mathbf{r}) = 0 \tag{2.1}$$

where $x(t,\mathbf{r})$ is a function representing the sound pressure at a point in time and space,

$$\mathbf{r} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{2.2}$$

and ∇^2 is the Laplacian operator. The speed of propagation, c, is dependent upon atmospheric conditions; the most important factor being temperature, with a 3% error affecting the performance of an array. In a normal room c is about 342m/sec. c varies with the square root of the absolute temperature yielding, $c = 331.4\sqrt{1+\theta/273}\ ms^{-1}$, where θ is in degrees centigrade. The wave equation, Equation (2.1), is known as the governing equation for a wide range of propagating waves, including electromagnetic waves. Using the method of separation of variables, the solution to the differential wave equation can be derived. The solution for a monochromatic plane wave is given as [12]

$$x(t,\mathbf{r}) = Ae^{j(\omega t - \mathbf{k}\cdot\mathbf{r})} \tag{2.3}$$

where A is the wave amplitude, $\omega = 2\pi f$ is the frequency in radians per second, and the *wavenumber vector* $\mathbf{k}$ indicates the speed and direction of the wave propagation, and is given by

$$\mathbf{k} = \frac{2\pi}{\lambda}\begin{bmatrix}\sin\theta\cos\phi & \sin\theta\sin\phi & \cos\theta\end{bmatrix} \tag{2.4}$$

where the wavelength λ is related to c by the simple equation $\lambda = c/f$ and the angles θ and ϕ are as shown in Figure 2-2. Alternatively, the solution for a spherical wave can be derived as [12]

$$x(t,\mathbf{r}) = -\frac{A}{4\pi r} e^{j(\omega t - kr)} \tag{2.5}$$

where $r = |\mathbf{r}|$ is the radial distance from the source, and k is the *scalar wavenumber*, given by $2\pi/\lambda$. The spherical wave solution shows that the signal amplitude decays at a rate proportional to the distance from the source. This dependence has important implications for array processing algorithms when the source is in the near-field, as will be discussed later. While sound waves are spherical in nature, they may be considered as plane waves at a sufficient distance from the source, and this approximation is often used to simplify mathematical analysis by introducing linearity to the wave solution [6].

The plane wave solution in Equation (2.3) is expressed in terms of two variables, time and space. Due to the well defined propagation of the signal, these two variables can be linked by a simple relation. Thus the solution can be expressed as a function of a single variable. If the solution of the plane wave is formulated as

$$x(t,\mathbf{r}) = Ae^{j\omega(t-\beta\cdot\mathbf{r})} \tag{2.6}$$

where $\beta = \mathbf{k}/\omega$, and we define a new variable u such that $u = t - \beta\cdot\mathbf{r}$, then the solution can be expressed as

$$x(u) = Ae^{j\omega u} \tag{2.7}$$

For spherical waves, with the substitution $u = t - r/c$, we have a similar expression

$$x(u) = -\frac{A}{4\pi r} e^{j\omega u} \tag{2.8}$$

Due to the linearity of the wave equation, the monochromatic solution can be expanded to the more general polychromatic case by considering the solution as a sum or integral of such complex exponentials [6]. Fourier theory states that any function with a convergent Fourier integral can be expressed as a weighted superposition of complex exponentials. From this it can be concluded that *any signal*

with a valid Fourier transform, irrespective of its shape, satisfies the wave equation [6].

In this section we see that propagating acoustic signals can be expressed as functions of a single variable, with time and space linked by a simple relationship. In addition, the information in the signal is preserved as it propagates. This means that, for a band-limited signal, the signal can be reconstructed over all space and time using the following methods:

1. *temporally* sampling the signal at a given *location in space*, or
2. *spatially* sampling the signal at a given *instant of time*.

The latter is the basis for all aperture and sensor array processing techniques. Other implications from the above wave propagation analysis that are important for array processing applications are, according to [13]:

1. The speed of propagation is dependent on the medium. For the specific case of acoustic waves in air, the speed of propagation is approximately $c = 331.4\sqrt{1+\theta/273}\ \ ms^{-1}$, with θ in degrees centigrade.
2. Waves generally propagate from their source as spherical waves, with the amplitude decaying at a rate proportional to the distance from the source.
3. The superposition principle applies to propagating wave signals, allowing multiple waves to occur without interaction. To separate these signals, algorithms must be developed to distinguish between different signals based on their temporal and spatial characteristics.

2.2 Apertures

The term *aperture* is used to describe a spatial region that transmits or receives propagating waves. A transmitting aperture is referred to as an *active aperture*, while a receiving aperture is called a *passive aperture*. In acoustics, an aperture is an electro-acoustic transducer that converts acoustic signals into electrical signals (microphone), or vice-versa (loudspeaker).

2.2.1 Aperture function

Consider a general receiving aperture of volume V where a signal $x(t,\mathbf{r})$ is received at time t and spatial location $\mathbf{r}$. Treating the infinitesimal volume dV at $\mathbf{r}$ as a linear filter having impulse response $a(t,\mathbf{r})$, the received signal is given by the convolution [12]

$$x_R(t,\mathbf{r}) = \int_{-\infty}^{\infty} x(\tau,\mathbf{r})a(t-\tau,\mathbf{r})d\tau \tag{2.9}$$

or by taking the Fourier transform,

$$X_R(f,\mathbf{r}) = X(f,\mathbf{r})A(f,\mathbf{r}) \tag{2.10}$$

$A(f,\mathbf{r})$ is known as the *aperture function* or the *sensitivity function* and it defines the response as a function of spatial position along the aperture.

2.2.2 Directivity pattern

The response of a receiving aperture is essentially directional because the amount of signal seen by the aperture varies with the direction of arrival of the signal. This principle is illustrated in Figure 2-1 [14] for planar waves being received by a linear aperture.

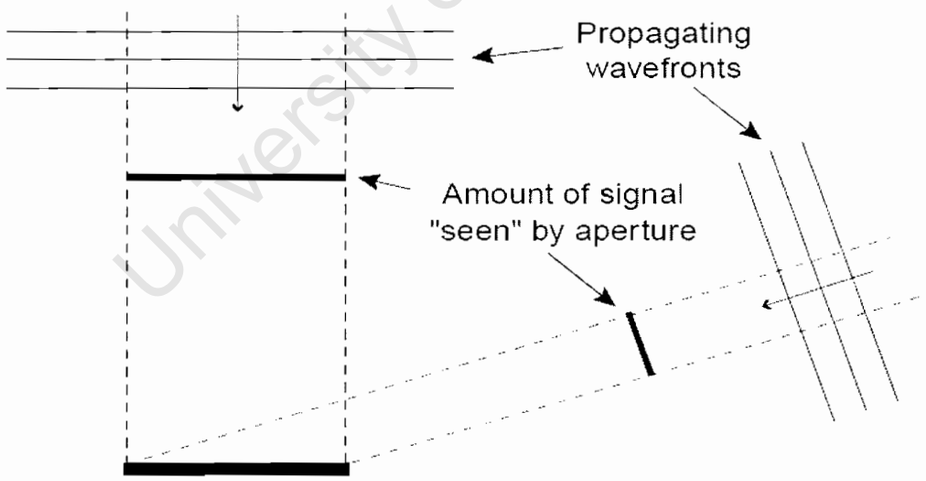

Figure 2-1: Signal received by linear aperture [14]

The aperture response as a function of frequency and direction of arrival is known as the aperture *directivity pattern* or *beam pattern*. By manipulating the solution to the wave equation discussed in section 2.1, the directivity pattern can be shown to be related to the aperture function by a Fourier transform relationship [12]. The far-field directivity pattern of a receiving aperture with aperture function A_R, is given by

$$D_R(f,\boldsymbol{\alpha}) = F_{\mathbf{r}}\{A_R(f,\mathbf{r})\} = \int_{-\infty}^{\infty} A_R(f,\mathbf{r}) e^{j2\pi\boldsymbol{\alpha}\cdot\mathbf{r}} d\mathbf{r} \tag{2.11}$$

where $F_{\mathbf{r}}\{\cdot\}$ denotes the three dimensional Fourier transform,

$$r = \begin{bmatrix} x_a \\ y_a \\ z_a \end{bmatrix} \tag{2.12}$$

is the spatial location of a point along the aperture, and

$$\boldsymbol{\alpha} = f\boldsymbol{\beta} = \frac{1}{\lambda}\begin{bmatrix}\sin\theta\cos\phi & \sin\theta\sin\phi & \cos\theta\end{bmatrix} \tag{2.13}$$

is the directional vector of the wave, where the angles θ and ϕ are as shown in Figure 2-2 [6]. Note that the frequency dependence in the above equations is implicit in the wavelength term as $\lambda = c/f$

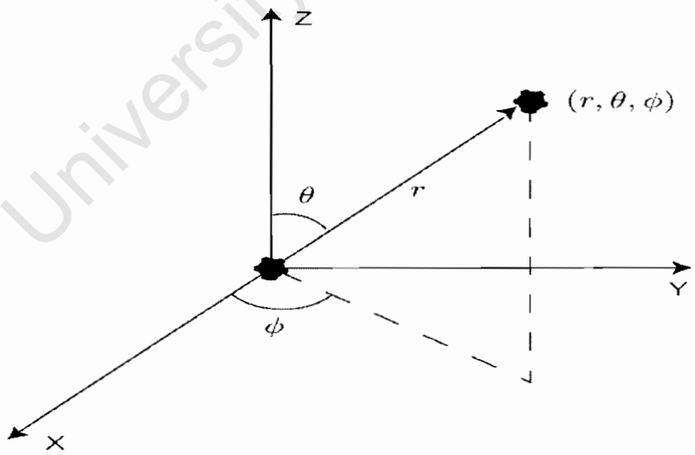

Figure 2-2: Spherical coordinate system [6]

2.2.3 Linear apertures

To investigate some of the properties of the aperture directivity pattern, it is necessary to simplify the above equation by considering a linear aperture of length L along the x-axis, as shown in Figure 2-3.

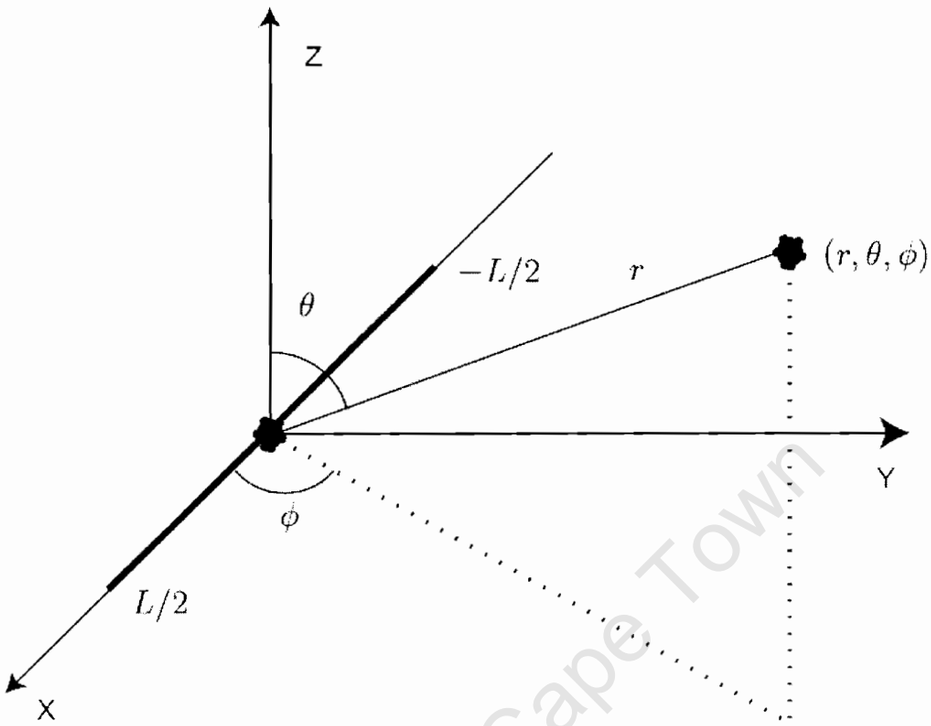

Figure 2-3: Continuous linear aperture running from *–L/2 to L/2* [6]

In this case

$$r = \begin{bmatrix} x_a \\ 0 \\ 0 \end{bmatrix} \tag{2.14}$$

and the directivity pattern simplifies to

$$D_R(f, \alpha_x) = \int_{-L/2}^{L/2} A_R(f, x_a) e^{j2\pi\alpha_x x_a} dx_a \tag{2.15}$$

where

$$\alpha_x = \frac{\sin\theta\cos\phi}{\lambda} \tag{2.16}$$

if the equation is written as a function of the angles θ and ϕ the obtained result is

$$D_R(f,\theta,\phi) = \int_{-L/2}^{L/2} A_R(f,x_a) e^{j\frac{2\pi}{\lambda}\sin\theta\cos\alpha x_a dx_a} \tag{2.17}$$

These expressions are only valid for the case of *far-field* sources because they have been developed for plane waves. According to [15] a plane wave can be considered to come from the far-field of the aperture if

$$|r| > \frac{2L^2}{\lambda} \tag{2.18}$$

where L is the length of the linear aperture along the x-axis and λ is the wavelength. The far-field assumption serves to simplify the discussion of aperture properties. Details of near-field sources will be discussed later when reviewing discrete linear sensor arrays.

Consider the case of a linear aperture with uniform, frequency-independent aperture function. The aperture function may be written as

$$A_R(x_a) = rect(x_a/L) \tag{2.19}$$

where

$$rect(x_a/L) \cong \begin{cases} 1 & |x| \le L/2 \\ 0 & |x| > L/2 \end{cases} \tag{2.20}$$

The resulting directivity pattern is given by

$$D_R(f,\alpha_x) = F\{rect(x_a/L)\} \tag{2.21}$$

which has the solution

$$D_R(f,\alpha_x) = L\sin c(\alpha_x L) \tag{2.22}$$

where

$$\sin c(x) \cong \frac{\sin(x)}{x} \tag{2.23}$$

Figure 2-4 shows plots of the uniform aperture function and a corresponding directivity pattern.

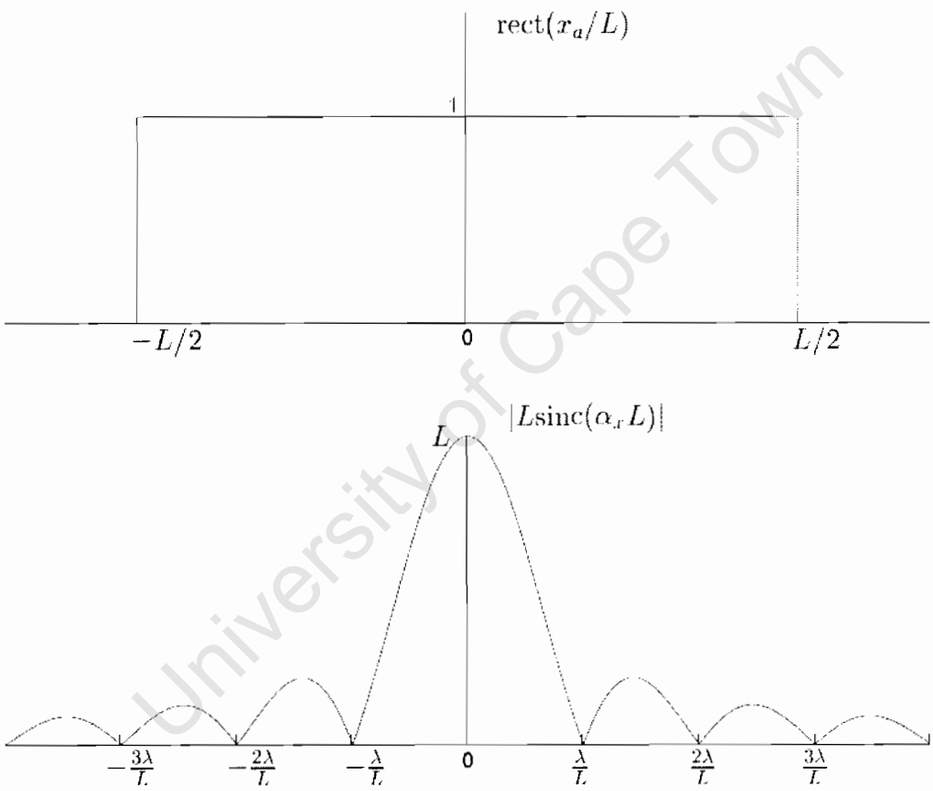

Figure 2-4: Uniform aperture function and corresponding directivity pattern [6]

From Figure 2-4 we see that zeros in the directivity pattern are located at $\alpha_x = m\lambda/L$, where m is an integer. The area in the range $-\lambda/L \le \alpha_x \le \lambda/L$ is referred to as the *main lobe* and its extent is termed the *beam width*. Thus it is seen that the beam width of a linear aperture is given by $2\lambda/L$, or in terms of frequency, $2c/fL$. We note that the beam width is inversely proportional to the product fL and so for a fixed aperture length, the beam width will decrease with increasing frequency.

The relative differences in array response over varying angles of signal arrival can be highlighted by considering the *normalized directivity pattern* of an aperture. The *sinc* function is bound by $-1 \le \sin c(x) \le 1$, and therefore the maximum possible value of the directivity pattern is $D_{\max} = L$, and the normalized directivity pattern is given as

$$D_N(f, \alpha_x) = \frac{D_R(f, \alpha_x)}{D_{\max}} = \sin c(\alpha_x L) \tag{2.24}$$

or in terms of the angles θ and ϕ

$$D_N(f, \theta, \phi) = \sin c(\frac{L}{\lambda} \sin\theta \cos\phi) \tag{2.25}$$

The properties of the aperture response are commonly examined using a polar plot of the horizontal directivity pattern over angle ϕ, given by

$$D_N(f, \frac{\pi}{2}, \phi) = \sin c(\frac{L}{\lambda} \cos\phi) \tag{2.26}$$

Polar plots of the horizontal directivity pattern for different values of L/λ are shown in Figure 2-5. The figures (a) to (d) demonstrate the beam width's dependence on the ratio L/λ.

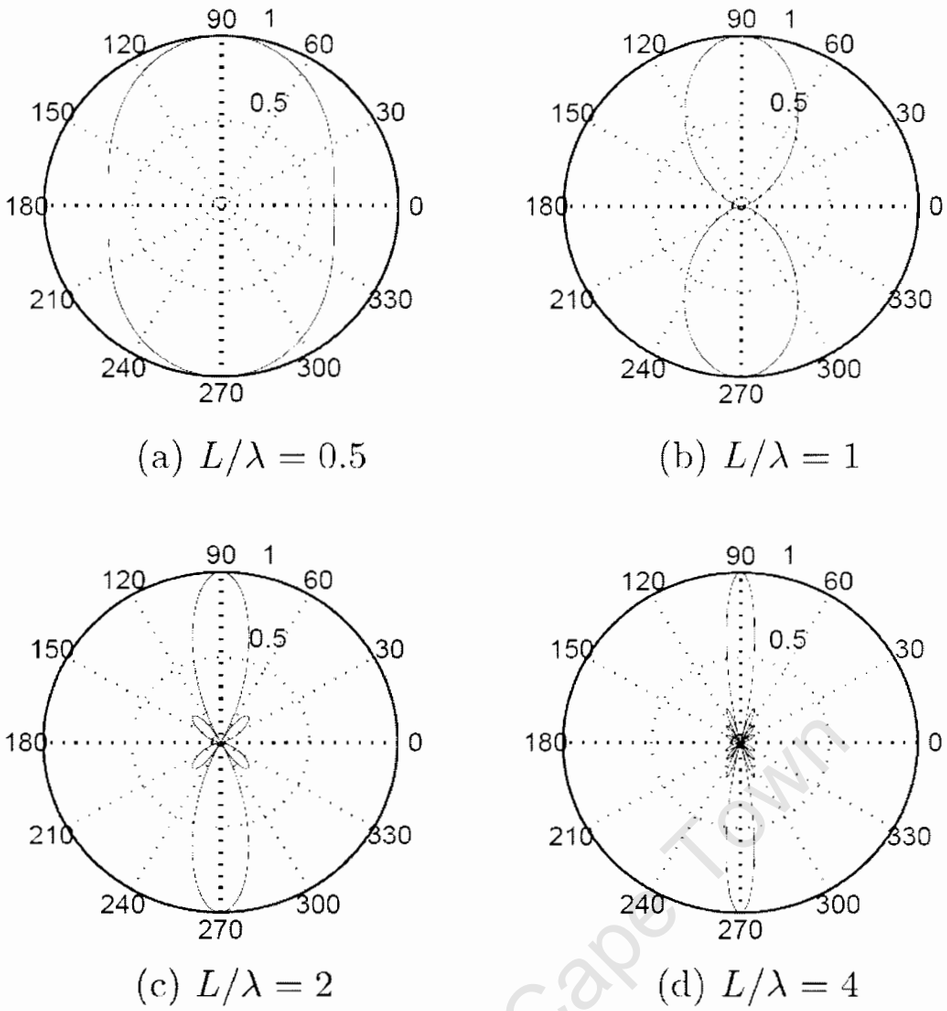

Figure 2-5: Directivity pattern polar plot [6]

2.3 Discrete sensor arrays

A discrete sensor array can be considered to be a sampled version of a continuous aperture. In this case, the aperture would only be excited at a finite number of discrete points. If each element is itself considered to be a continuous aperture, then the overall response of the array can be determined as the superposition of each individual sensor response [6].

2.3.1 Linear sensor arrays

For simplicity, consider the case of a linear array having an odd number of elements as shown in Figure 2-6.

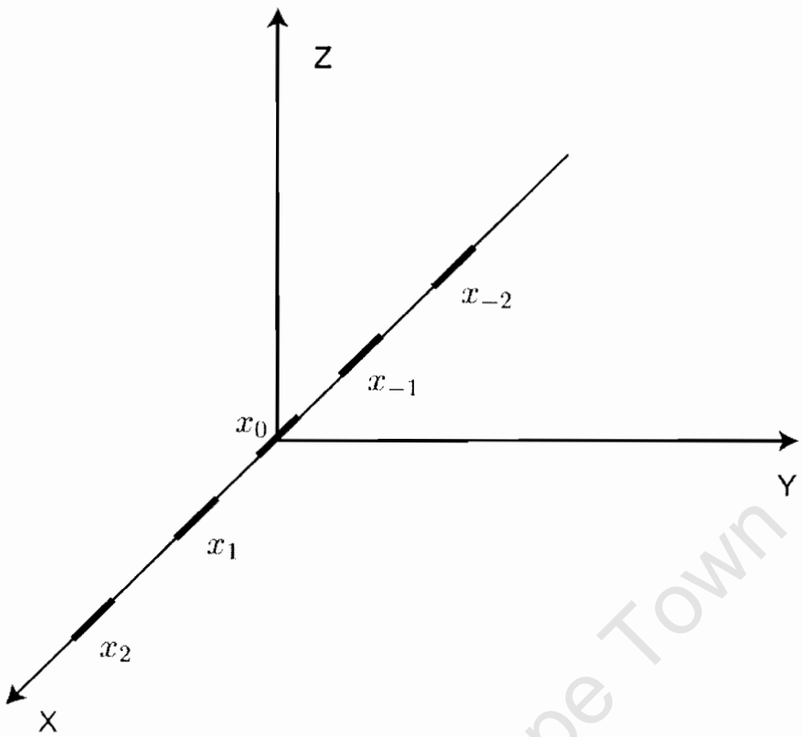

Figure 2-6: Discrete sensor array [6]

Using the superposition principle we can express the complex frequency response of the array for the general case where each element has a different complex frequency response $e_n(f,x)$, as

$$A(f,x_a)=\sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} w_n(f)e_n(f,x_a-x_n) \tag{2.27}$$

where $w_n(f)$ is the complex weight for element n, $e_n(f,x)$ is its complex frequency response or *element function*, and x_n is its spatial position on the x-axis. In the case

where all the elements have identical frequency response (i.e. $e_n(f,x) = e(f,x)$ for all values of n), the aperture function can be simplified to

$$A(f,x_a) = \sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} w_n(f)\delta(x_a - x_n) \tag{2.28}$$

Substituting this discrete aperture response function into the directivity pattern equation, Equation (2.15), we obtain the far-field directivity pattern as

$$D(f,\alpha_x) = \sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} w_n(f)e^{j2\pi\alpha_x x_n} \tag{2.29}$$

Equation (2.29) is the far-field directivity pattern equation for a linear array of N identical sensors, with arbitrary *inter-element* spacing. For the case where all the elements are equally spaced by d meters, the directivity pattern becomes

$$D(f,\alpha_x) = \sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} w_n(f)e^{j2\pi\alpha_x nd} \tag{2.30}$$

and if only the horizontal directivity pattern is considered, the following equation is obtained

$$D(f,\phi) = \sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} w_n(f)e^{j\frac{2\pi}{\lambda}nd\cos\phi} \tag{2.31}$$

or, making the frequency dependence explicit the resultant equation becomes

$$D(f,\phi) = \sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} w_n(f)e^{j\frac{2\pi f}{c}nd\cos\phi} \tag{2.32}$$

This equation gives the directivity pattern for a linear, equally spaced array of identical sensors. From this equation it is evident that the directivity pattern is dependent upon

1. The number of array elements N
2. The inter-element spacing d, and
3. The frequency f.

The *effective length* of a sensor array is the length of the continuous aperture which it samples, and is given by $L = Nd$. The actual *physical length* of the array is the distance between the first and last sensors that is $d(N-1)$. Varying the three variables mentioned above independently and plotting directivity patterns exhibits characteristics of a linear, equally spaced sensor array. Figure 2-7 is a plot of the directivity pattern for the case where the number of elements in the array N, is varied and the effective length and frequency are kept constant.

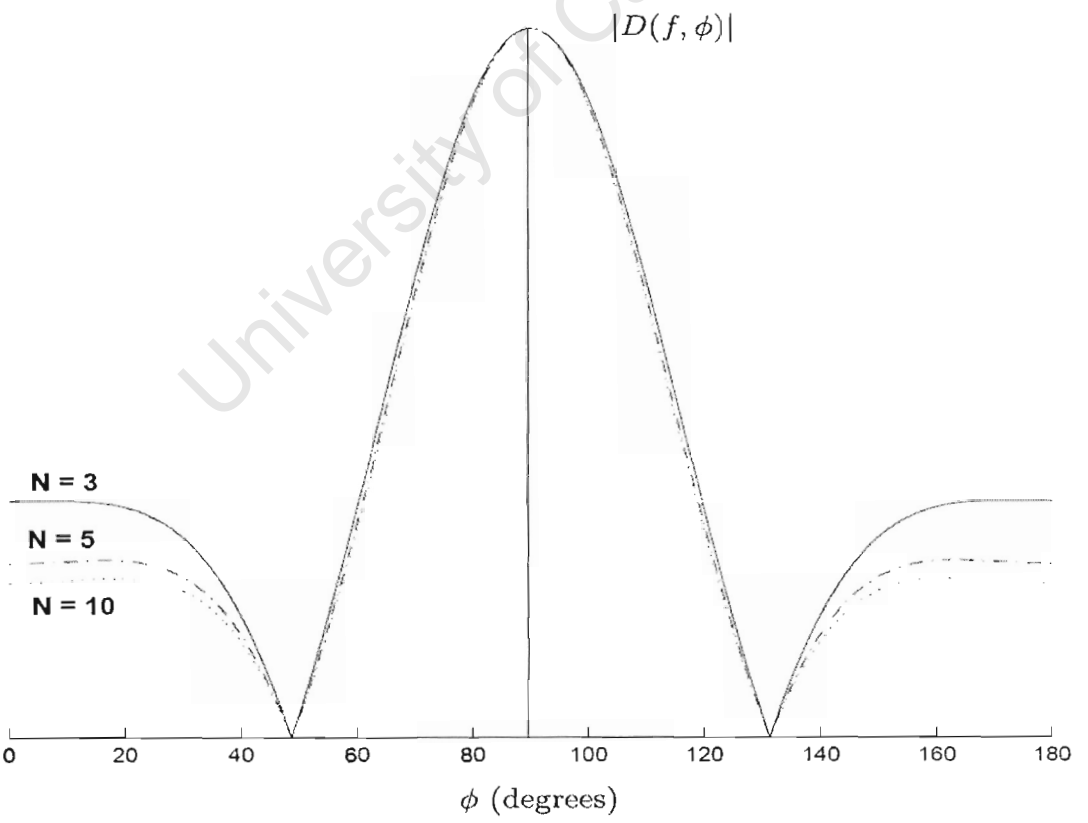

Figure 2-7: Directivity pattern for varying number of sensors (f=1 kHz, L=0.5m) [6]

It is noted that the sidelobe level decreases with increasing sampling frequency – that is, the more sensors used, the lower the sidelobe level. Figure 2-8 shows the directivity pattern for the case where the effective length L, of the array is varied and number of elements and frequency are kept constant. The plot shows that the beam width decreases as the effective length (and thus the inter-element spacing) increases. The beam width is actually inversely proportional to the product fL. Given that $L = Nd$ and that N is fixed in this case, it can be seen that to vary the beam width we must vary fd. However, it is more common to require that the beam width remain constant, in which case fd must be kept relatively constant. It can thus be seen that for a given frequency, two important characteristics of the array directivity pattern, namely the beam width and the sidelobe level, are directly determined by the inter-element spacing and the number of sensors respectively.

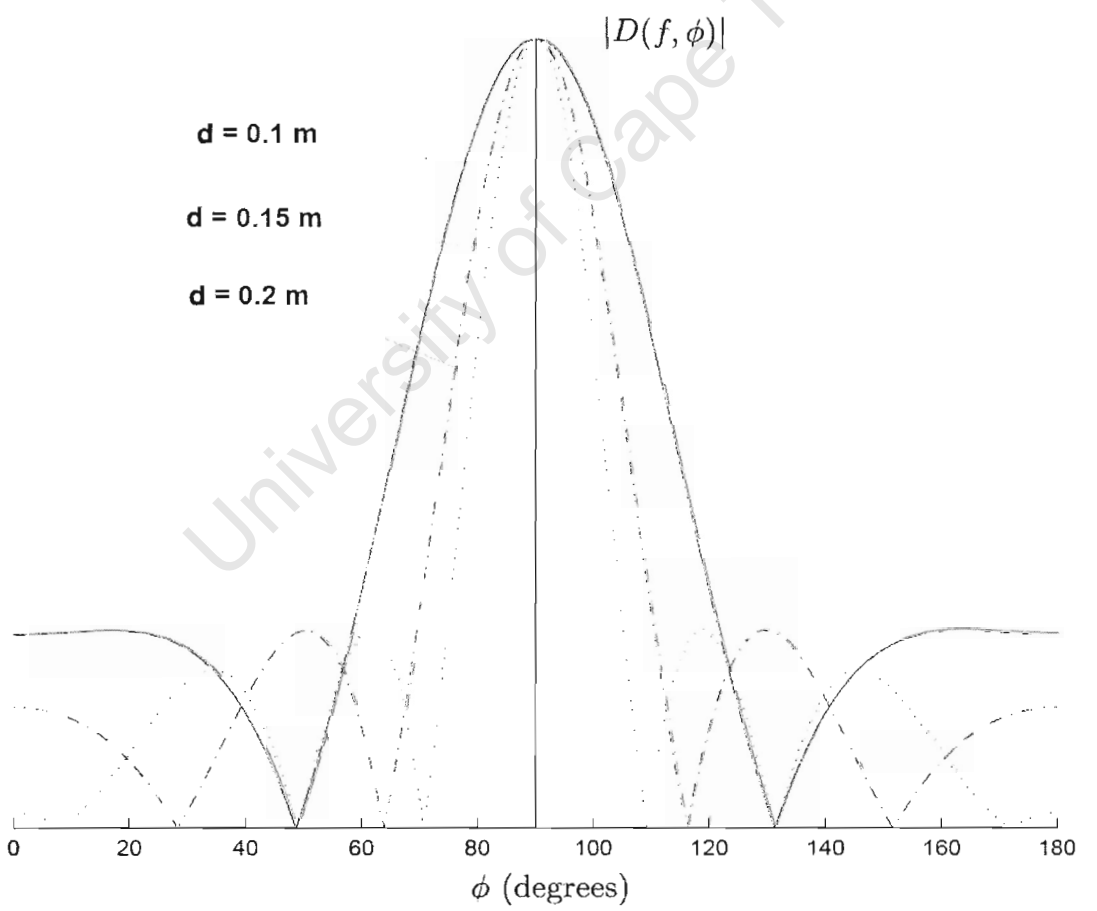

Figure 2-8: Directivity pattern for varying effective array length (f=1 kHz, N=5) [6]

Figure 2-9 shows the effect of varying the frequency while keeping the effective length and the number of array elements constant.

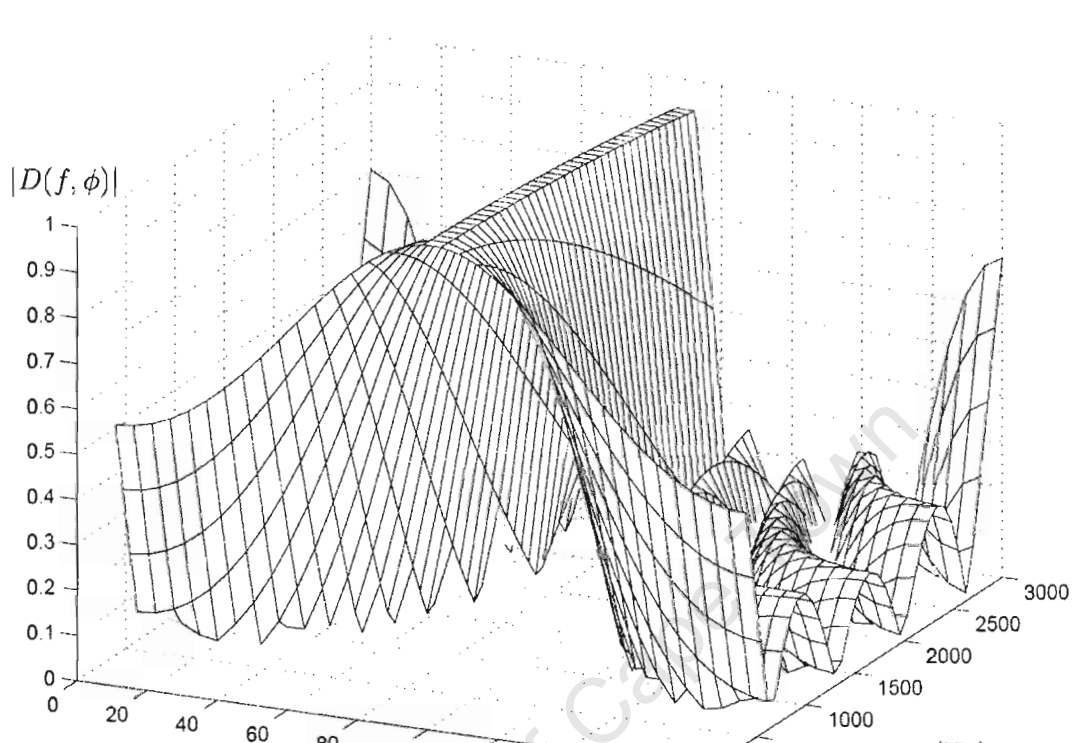

Figure 2-9: Directivity pattern for the range 400 to 3000Hz (N=5, d=0.1m) [6]

For a given array configuration, the beam width will vary as a function of frequency. As the frequency increases, the beam width will decrease. Figure 2-9 is a plot of the horizontal directivity pattern where the frequency is varied over the range $400Hz \leq f \leq 3000Hz$.

In temporal sampling the principle of Nyquist frequency has to be taken into consideration. Nyquist frequency is the minimum sampling frequency required in order to avoid aliasing in the sampled signal [16]. The temporal sampling theory states that a signal must be sampled at a rate f_s (period T_s) such that

$$f_s = \frac{1}{T_s} \geq 2f_{\max} \tag{2.33}$$

where f_{max} is the maximum frequency component in the signal's frequency spectrum. Similarly, spatial sampling has the requirement that

$$f_{x_s} = \frac{1}{d} \geq 2 f_{x_{max}} \tag{2.34}$$

where f_{x_s} is the spatial sampling frequency in samples per meter and $f_{x_{max}}$ is the highest spatial frequency component in the angular spectrum of the signal. The spatial sampling frequency along the x-axis is given by

$$f_{x_s} = \frac{\sin\theta\cos\phi}{\lambda} \tag{2.35}$$

The maximum of this ratio occurs when the numerator attains its maximum and the denominator attains its minimum which leads to

$$f_{max} = \frac{1}{\lambda_{min}} \tag{2.36}$$

and consequently the requirement that

$$d < \frac{\lambda_{min}}{2} \tag{2.37}$$

where λ_{min} is the minimum wavelength in the signal of interest. This equation is known as the *spatial sampling theorem*, and must be adhered to in order to avoid *spatial aliasing* in the directivity pattern of a sensor array.

2.3.2 Near-field sources

So far, only the case of far-field sources has been considered. Recall from Section 2.2.3 that, for a linear aperture, a wave source may be considered to come from the far-field of the aperture if

$$|r| > \frac{2L^2}{\lambda} \tag{2.38}$$

Under this assumption, the curvature of the wavefront can be neglected, that is, the wavefronts arriving at the aperture can be considered as plane waves. For many practical applications particularly in speech recognition, the above condition is not met and the source is said to be located in the *near-field* of the array. As the derivation of the equivalent near-field expressions for the general continuous and discrete directivity patterns is quite involving, it is sufficient to consider only the derivation of the expression for the horizontal directivity pattern for a linear sensor array.

Considering the arrival of planar wavefronts on different elements in a sensor array it can be seen from Figure 2-10 that the actual distance traveled by the wave between adjacent sensors is given by

$$d' = d\cos\phi \tag{2.39}$$

More generally, the distance traveled by the wave between the reference sensor, $n = 0$, and the n^{th} sensor is given by

$$d' = nd\cos\phi \tag{2.40}$$

On the other hand, Figure 2-11 shows the arrival of spherical wavefronts on different elements in a sensor array. It is observed from the diagram that the actual distance traveled by the wave between the two sensors is given by

$$d' = d_1(r,\phi) - d_0(r,\phi) \tag{2.41}$$

and in general

$$d' = d_n(r,\phi) - d_0(r,\phi) \tag{2.42}$$

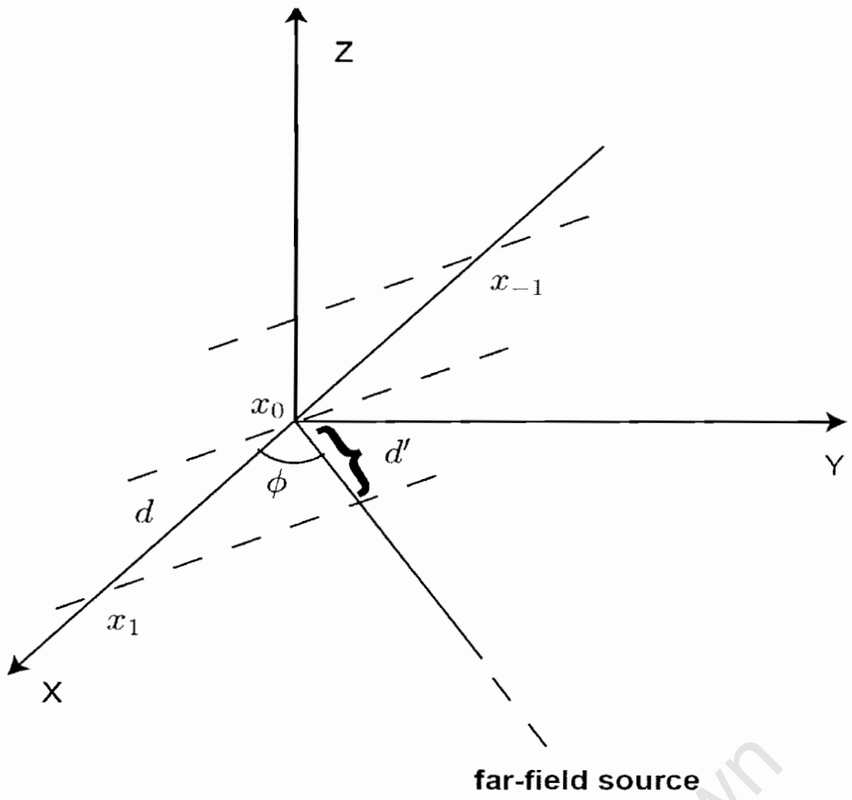

Figure 2-10: Arrival of wavefronts from far-field source [6]

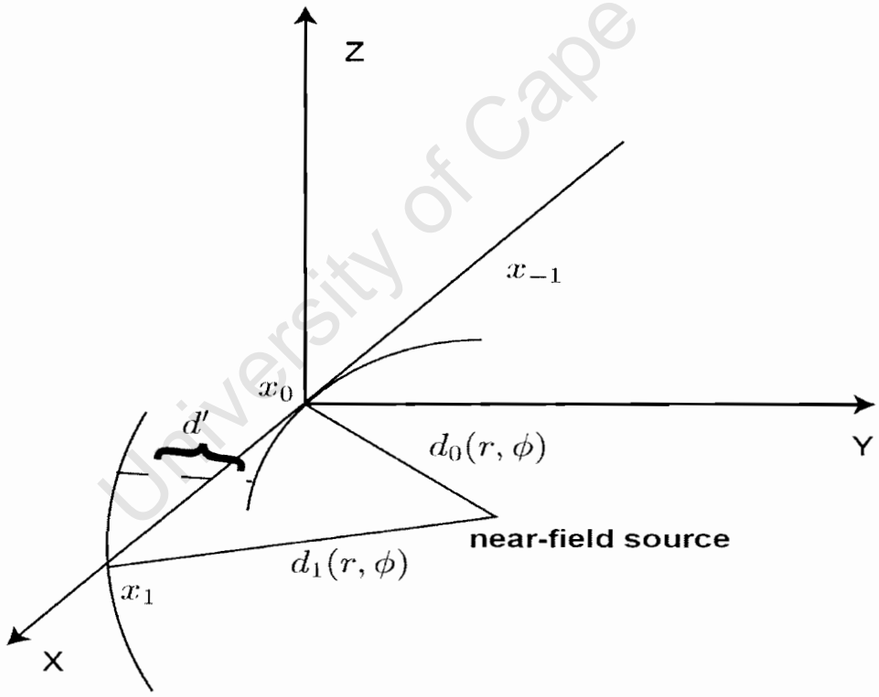

Figure 2-11: Arrival of wavefronts from near-field source [6]

In the above equation, $d_n(r,\phi)$ is the distance from the source to the n^{th} sensor as a function of the spherical coordinates of the source (in the horizontal plane) with respect to the reference sensor. Using trigonometric relations, this distance can be given by [17]

$$d_n(r,\phi) = [r^2 + 2r(x_n - x_0)\cos\phi + (x_n - x_0)^2]^{\frac{1}{2}} \tag{2.43}$$

which, in the case of an equally spaced array, reduces to

$$d_n(r,\phi) = [r^2 + 2rnd\cos\phi + (nd)^2]^{\frac{1}{2}} \tag{2.44}$$

Looking back at the far-field horizontal directivity pattern equation for a linear sensor array

$$D(f,\phi) = \sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} w_n(f) e^{j\frac{2\pi}{\lambda}nd\cos\phi} \tag{2.45}$$

it can be noted that the exponential contains the term $nd\cos\phi$, and it has been shown that this corresponds to the distance traveled by the propagating wave between the reference sensor and the n^{th} sensor. Substituting it in the equivalent expression for the near-field case the resulting equation is

$$D'(f,\phi) = \sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} w_n(f) e^{j\frac{2\pi}{\lambda}(d_n(r,\phi)-d_0(r,\phi))} \tag{2.46}$$

As discussed earlier in Section 2.1, recalling that the amplitude for spherical acoustic waves decays at a rate proportion to the distance traveled, for far-field sources the amplitude differences between sensors can be considered to be negligible but may be significant for near-field sources. Incorporating the amplitude dependency into the expression and normalizing to give the reference sensor unity amplitude, the following directivity pattern equation for near-field sources is obtained

$$D_{nf}(f,\phi) = \sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} \frac{d_0(r,\phi)}{d_n(r,\phi)} w_n(f) e^{j\frac{2\pi}{\lambda}(d_n(r,\phi)-d_0(r,\phi))} \tag{2.47}$$

Figure 2-12 is a plot showing the horizontal directivity pattern for both a far-field source and a near-field source for the same sensor array for r = 1m.

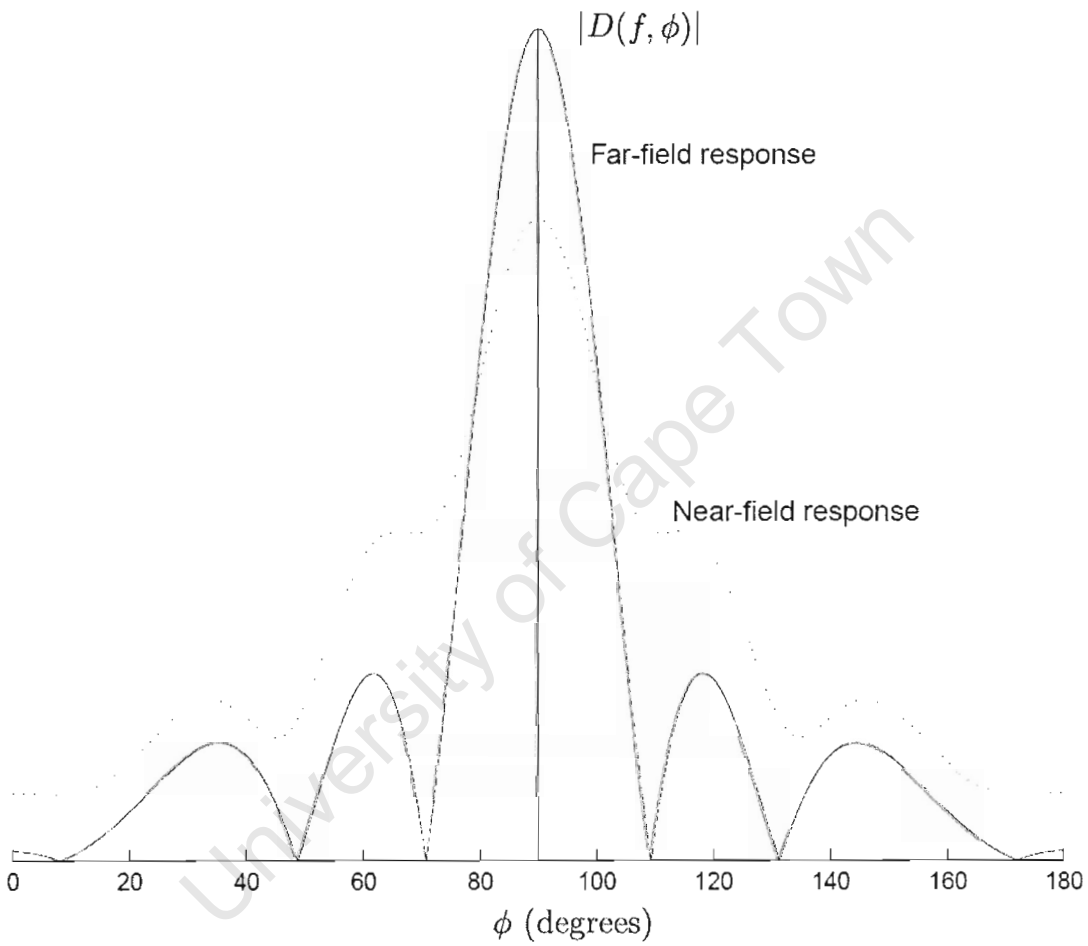

Figure 2-12: Directivity pattern for far-field and near-field source (f=1kHz, N=10, d=0.1m) [6]

If a microphone array is desired to operate in the near-field, the near-field directivity pattern can be made to match the corresponding far-field directivity pattern by compensating the frequency dependent sensor weights $w_n(f)$. If the far-field weights are replaced with the near-field compensated weights

$$D'(f,\phi) = \frac{d_n(r,\phi)}{d_0(r,\phi)} e^{j\frac{2\pi}{\lambda}(d_0(r,\phi)-d_n(r,\phi)+nd\cos\phi)} w_n(f) \tag{2.48}$$

then the near-field directivity pattern will match the far-field directivity pattern obtained using the original weights. This procedure is called *near-filed compensation* and allows for the approximation of a desired far-field directivity pattern at a given point (r,ϕ), in the near-field [6].

2.4 Beamforming

Presently it is considered that the weights $w_n(f)$, in the far-field directivity pattern equation of a linear sensor array are equally weighted, that is

$$w_n(f) = \frac{1}{N} \tag{2.49}$$

In general, the complex weighting can be expressed in terms of its magnitude and phase components as

$$w_n(f) = a_n(f)e^{j\varphi_n(f)} \tag{2.50}$$

where $a_n(f)$ and $\varphi_n(f)$ are real, frequency dependent amplitude and phase weights respectively. By modifying the amplitude weights we can change the shape of the directivity pattern and by modifying the phase weights we can control the angular location of the response's main lobe. *Beamforming techniques* are algorithms for determining the complex sensor weights in order to implement a desired *shaping* and *steering* of the array directivity pattern [6]. These algorithms take advantage of the time differentials between incoming signals among the sensors in the array. This is due to the fact that a signal emitted from a source located at a particular position in space will arrive at each sensor at a specific time according to the relative positioning of the sensors in the array and the source.

In illustrating the concept of beam steering, we consider the case where the amplitude weights are set to unity, resulting in the directivity pattern

$$D(f,\phi) = \sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} e^{j(2\pi\alpha_x nd + \varphi_n(f))} \tag{2.51}$$

If we use the phase weights

$$\varphi_n(f) = -2\pi\alpha'_x nd \tag{2.52}$$

where

$$\alpha'_x = \frac{\sin\theta'\cos\phi'}{\lambda} \tag{2.53}$$

then the directivity pattern becomes

$$D'(f,\phi) = \sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} e^{j\frac{2\pi}{\lambda}nd(\alpha_x - \alpha'_x)} \tag{2.54}$$

which can be expressed as

$$D'(f,\phi) = D(f,\alpha_x - \alpha'_x) \tag{2.55}$$

The effect of such a phase weight on the beam pattern is to steer the main lobe of the beam pattern to the direction cosine $\alpha_x = \alpha'_x$, and thus to the directions $\theta = \theta'$ and $\phi = \phi'$. While the beam pattern remains unchanged apart from the shift along the α_x axis, when plotted as a function of angle, the beam shape will change as α_x is actually a function of $\sin\theta$ and $\cos\phi$. Figure 2-13 shows the horizontal directivity pattern shifted to $\phi' = 45^o$.

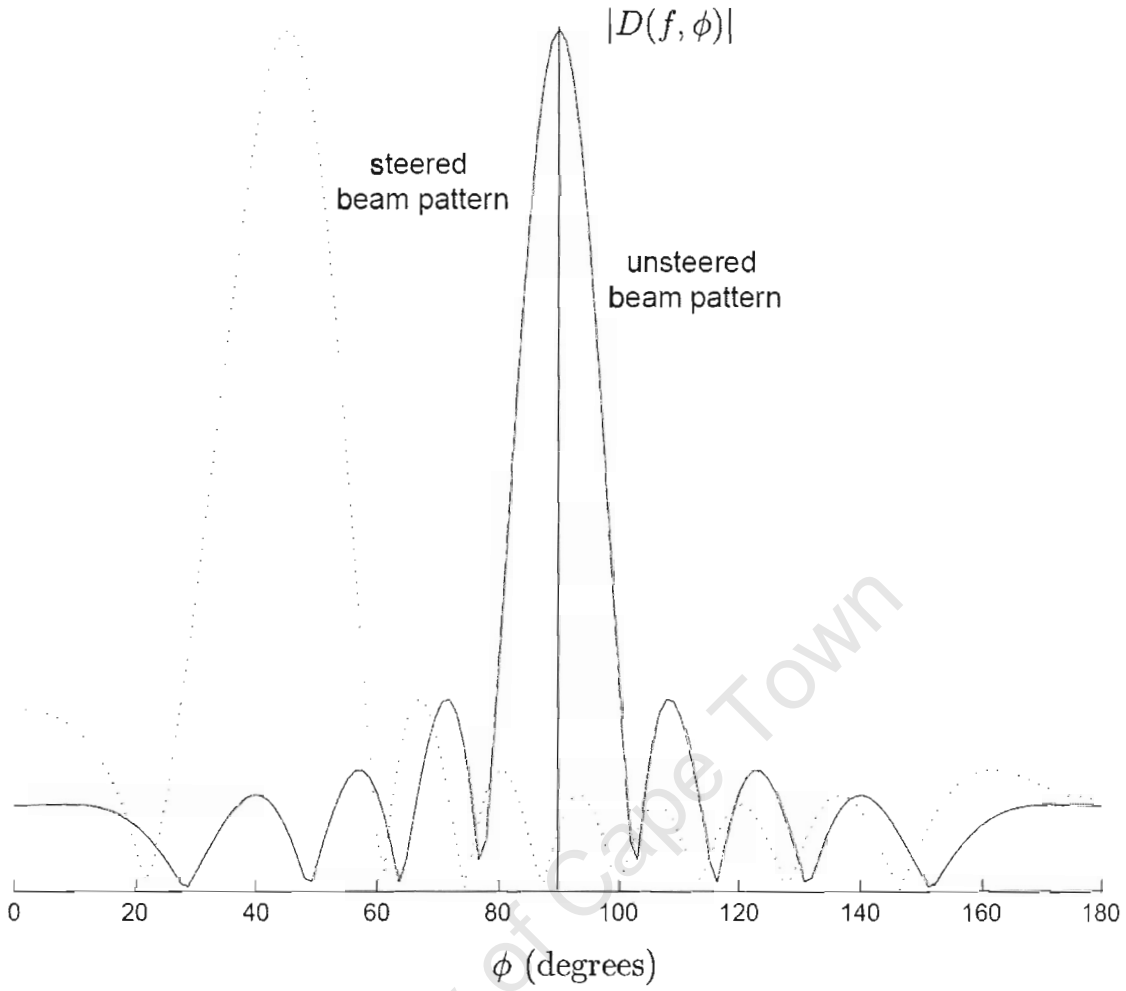

Figure 2-13: Unsteered and steered directivity patterns (φ' = 45 degrees, f = 1 kHz, N = 10, d = 0.15m) [18]

From Fourier transform theory, it is known that a negative phase shift in the frequency domain corresponds to a time delay in the time domain [16]. This implies that beam steering can effectively be implemented by applying time delays to the sensor inputs. Considering only the horizontal plane, the delay for the n^{th} sensor is given by

$$\begin{aligned}\tau_n &= \frac{\varphi_n}{2\pi f} \\ &= \frac{2\pi f n d \cos\phi'}{2\pi f c} \\ &= \frac{nd\cos\phi'}{c}\end{aligned} \tag{2.56}$$

which is seen to be equivalent to the time the plane wave takes to travel between the reference sensor and the n^{th} sensor. This is the principle of the simplest of all beamforming techniques, known as *delay-and-sum beamforming*.

Beamforming techniques are generally classified as being either *data-independent* or *data-dependent*. Data-independent or *fixed* beamformers are so named because their parameters are fixed during operation. On the other hand, data-dependent, or *adaptive* beamformers continuously update their parameters based on the received signals.

2.4.1 Delay-and-sum beamformer

The most fundamental of the beamforming algorithms is the *delay-and-sum* (DS) beamformer, which is a classical example of a data-independent beamformer [11]. The delay-and-sum beamforming technique adds captured signals from the array of sensors after aligning the signals in such a way that signal components originating from a desired direction are combined coherently, while signals originating from other directions are combined in an incoherent fashion.

Given a signal of interest in a certain location in space, the signal will arrive at the sensors in an array at times determined by each sensor's location. For a linear, equally spaced array these time differentials are as given previously in Equation (2.56). Once the time difference of each sensor relative to the others is determined each sensor signal is delayed by τ_n seconds to align the signal of interest and then summed to give a single array output [19]. The delay-and-sum beamformer achieves an increased gain for the main lobe in the direction of the desired signal with signal enhancement and noise reduction provided by the constructive (in phase) interference of the desired propagating wave and the destructive (out of phase) interference of the waves from all other directions. The signal gain over the undesired noise increases as a function of the number of sensors in the array [2]. Expressing the array output as the sum of the weighted channels, the following equation is obtained in the time domain

$$y(t) = \frac{1}{N}\sum_{n=1}^{N} x_n(t - \tau_n) \tag{2.57}$$

where τ_n represents the time delay for the n^{th} sensor. Many beamforming techniques that exist combine the conventional delay-and-sum beamformer with channel filters to implement a desired shaping of the beam pattern.

2.4.2 Filter-and-sum beamformer

While the delay-and-sum beamformer is easy to understand, it offers minimal noise reduction and requires a large number of microphones to improve SNR [11]. It belongs to a more general class of beamformers known as *filter-and-sum beamformers,* where both the amplitude and phase weights are frequency dependent. In practice, most beamformers are in the class of the filter-and-sum beamformer. Filter-and-sum beamformers apply filters to the array signals as well as time alignment. The derivation of the filters in the filter-and-sum beamformers is what distinguishes one from the other [4]. Both beamforming concepts, delay-and-sum and filter-and-sum, were first developed on the basis of the far-field assumption, but may also be extended to near-field beamforming [20]. The output of a filter-and-sum beamformer is given as

$$y(f) = \sum_{n=1}^{N} w_n(f) x_n(f) \tag{2.58}$$

It is often convenient to use matrix algebra to simplify the notation when describing microphone array techniques. It is assumed that speech, $\boldsymbol{s}$ and affecting noise, $\boldsymbol{n}$ are statistically uncorrelated, and that noise is linearly added to speech: $\mathbf{x} = \mathbf{s} + \mathbf{n}$, where, for example, $\mathbf{x}$ is the output from the N channels of the microphone array. The above equation can now be rewritten using matrix notation as

$$y(f) = w(f)^T x(f) \tag{2.59}$$

where the weight vector $\mathbf{w}(f)$ and the data vector $\mathbf{x}(f)$ are defined as

$$w(f) = [w_1(f) \cdots w_n(f) \cdots w_N(f)]^T \tag{2.60}$$

and

$$\mathbf{x}(f) = [x_1(f) \cdots x_n(f) \cdots x_N(f)]^T \tag{2.61}$$

where $(\cdots)^T$ denotes matrix transpose. Figure 2-14 shows the general structure of a filter-and-sum beamformer.

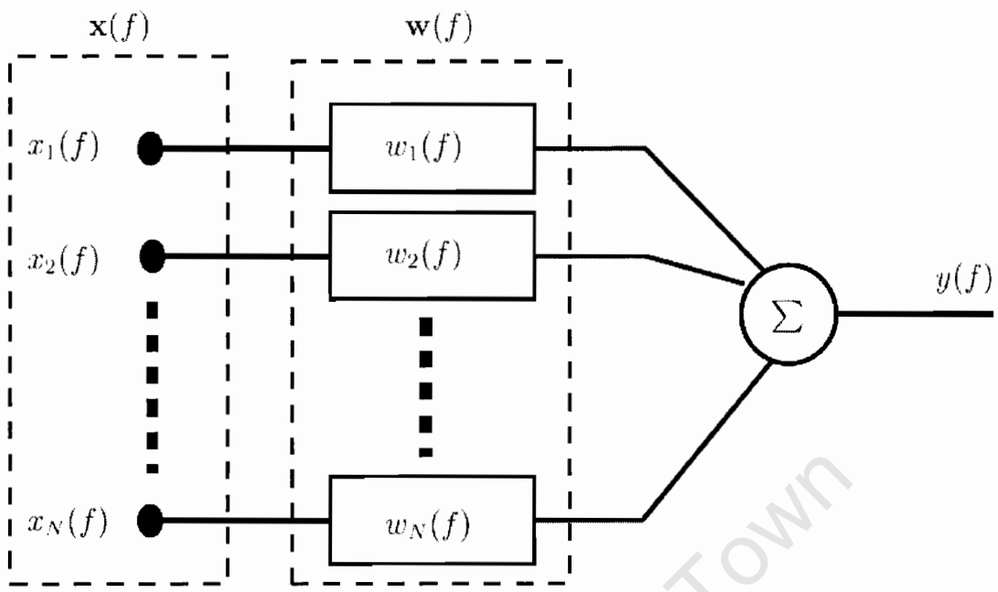

Figure 2-14: Filter-and-sum beamformer structure [6]

2.4.3 Generalized sidelobe canceller (GSC)

One limitation of data independent beamforming techniques, such as the delay-and-sum and the filter-and-sum is their inability to adapt to changing noise conditions. Data-dependent beamforming techniques, such as the Generalized Sidelobe Canceller (GSC) [21], aim to counter this limitation by adaptively filtering incoming signals. This is done in order to pass the signal from the desired direction while rejecting noises coming from other directions. The GSC is an example of an efficient implementation of an adaptive beamformer that minimizes the *mean square error* (MSE) between a reference signal that is highly correlated to the desired signal, and the output signal [9, 21]. Here, the adaptive beamformer is separated into two main parallel processing paths. The first path implements a standard fixed beamformer steered toward the desired source. The second path is the adaptive part, which provides a set of filters that adaptively minimize the noise power in the output. The

desired signal is blocked from the second path by a *blocking matrix*, ensuring that the noise power is minimized. Such an adaptive beamforming technique succeeds in significantly reducing the noise level for coherent noise signals emanating from localized sources [21]. Due to the blocking matrix, the lower path output only contains noise signals. The overall system output is calculated as the difference of the upper and lower path outputs

$$y(f) = y_u(f) - y_a(f) \tag{2.62}$$

The GSC is a flexible structure due to the separation of the beamformer into a fixed and adaptive portion. In practice, the GSC can cause a degree of distortion to the desired signal due to what is termed as signal leakage. This occurs when the blocking matrix fails to remove the desired signal entirely from the lower noise canceling path. The block structure of the generalized sidelobe canceller is shown in Figure 2-15.

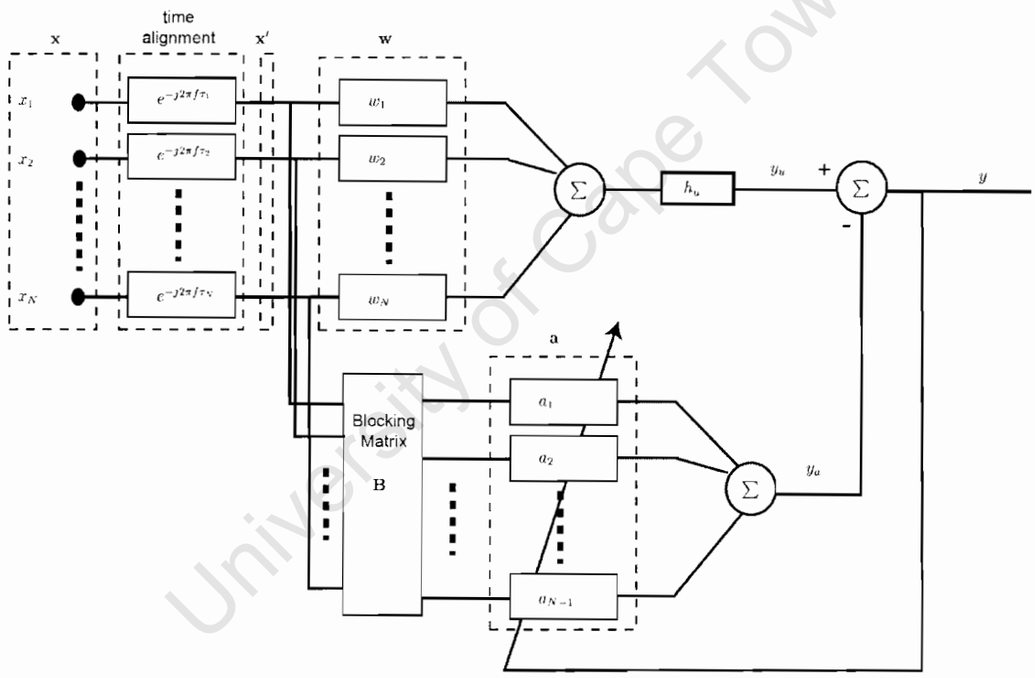

Figure 2-15: Generalized sidelobe canceller structure [6]

2.4.4 Post-filtering

In practice, the basic filter-and-sum beamformer rarely displays the level of improvement that theory promises. Further enhancement to improve the system performance is achieved through the addition of a post-filter at the output of the beamformer. Incorporating a post-filter with a beamformer allows for the use of knowledge obtained in spatial filtering to also allow for effective frequency filtering of the signal. In using both spatial and frequency domain enhancement, the use of signal information is maximized. The use of a post-filter with a filter-and-sum microphone array was thoroughly investigated by Marro [22] who demonstrated the mathematical interaction of the post-filter and the beamformer, and determined an optimal array structure for their combination. Figure 2-16 shows a diagram illustrating this system.

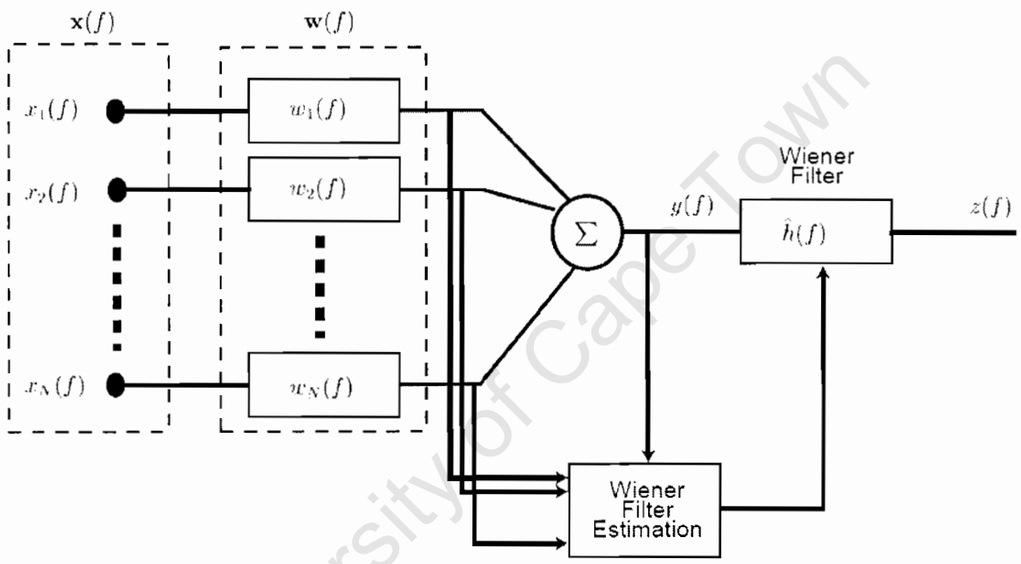

Figure 2-16: Filter-and-sum beamformer with post-filter [23]

2.5 Noise fields

There are three main categories of noise fields for microphone array applications. These categories are characterized by the degree of correlation between noise signals at different spatial locations. A commonly used measure of the correlation is *coherence*, which is defined as [24]

$$\Gamma_{ij}(f) \cong \frac{\Phi_{ij}(f)}{\sqrt{\Phi_{ii}(f)\Phi_{jj}(f)}} \tag{2.63}$$

where Φ_{ij} is the cross-spectral density between signals *i* and *j*. The coherence is essentially a normalized cross-spectral measure, as the magnitude squared can be seen to be bounded by $0 \le \left|\Gamma_{ij}(f)\right|^2 \le 1$.

2.5.1 Coherent noise fields

A *coherent noise field* is one in which noise signals propagate to the microphones directly from their sources without undergoing any form of reflection, dispersion or dissipation due to the acoustic environment [6]. In a coherent noise field, the noise signals on different microphones in an array are strongly correlated, and hence $\left|\Gamma_{ij}(f)\right|^2 \approx 1$. In practical applications, coherent noise fields occur in open air environments where there are no major obstacles and where wind or thermal turbulence effects are minimal.

2.5.2 Incoherent noise fields

In an *incoherent noise field*, also referred to as *spatial white noise,* the noise measured at any given spatial location is uncorrelated with the noise measured at all other locations, that is $\left|\Gamma_{ij}(f)\right|^2 \approx 0$. Such an ideal incoherent noise field is difficult to achieve and is rarely encountered in practical applications.

2.5.3 Diffuse noise field

In a *diffuse noise field*, noise of equal energy propagates in all directions simultaneously. Thus sensors in a diffuse noise field will receive noise signals that are poorly correlated, but have approximately the same energy. Many practical noise environments can be characterized by a diffuse noise field, such as office noise or car noise. The coherence between the noise at any two points in a diffuse noise field is a function of the distance between the sensors, and can be modeled as [25]

$$\Gamma_{ij}(f) = \sin c\left(\frac{2\pi f d_{ij}}{c}\right) \tag{2.64}$$

where d_{ij} is the distance between sensors i and j. It can be seen that the coherence approaches unity for closely spaced sensors and decreases sharply with increasing distance [6].

2.6 Summary

This chapter provides a comprehensive review of the fundamental techniques and basic concepts required to build a multi-sensor array. Continuous apertures and discrete sensor arrays are reviewed in detail. Some basic beamforming techniques were also discussed, as well as some of the common noise fields encountered in speech and microphone array processing. The following chapter provides some insight into basics of speaker recognition systems and the potential role of microphone arrays in these systems.

Chapter 3

3 Speaker Recognition and Microphone Arrays

Speaker recognition can be classified into *speaker identification* and *speaker verification*. It is the process of automatically determining an individual's identity on the basis of individual information included in their speech signals. Speaker recognition has a wide range of potential applications. This technology makes it possible to use a speaker's voice to verify their identity and control access to services such as voice dialing, database access services and security control. With the increased use of automated services for applications such as telephone banking, speaker recognition has the potential to become an important means of authentication over telephone networks. One of the several advantages of using speech to determine an individual's identity is that speech is the most natural means of interacting. Speaker recognition is generally regarded as being less intrusive to perform, as there is no need to place one's head in a specific position so that a system can scan your retina for example [26][3].

3.1 Overview on speaker recognition

Speaker recognition generally uses *pattern recognition* techniques. Pattern recognition is defined as "the study of how machines can observe the environment, learn to distinguish patterns of interest from their background, and make sound and reasonable decisions about the categories of the patterns" [27]. This generally

[3] Much of the information in this chapter was taken from this reference.

involves three aspects: (1) data acquisition and pre-processing; (2) data representation; and (3) decision making. In addition, at any time, a pattern recognition system is either in one of two modes of operation, that is, the *training mode* or the *testing mode*. In the training mode, the system "learns" the categories to which the input training patterns belong and, in the testing mode, patterns are classified according to their similarity to these categories [26]. Figure 3-1 shows a generic speaker recognition system.

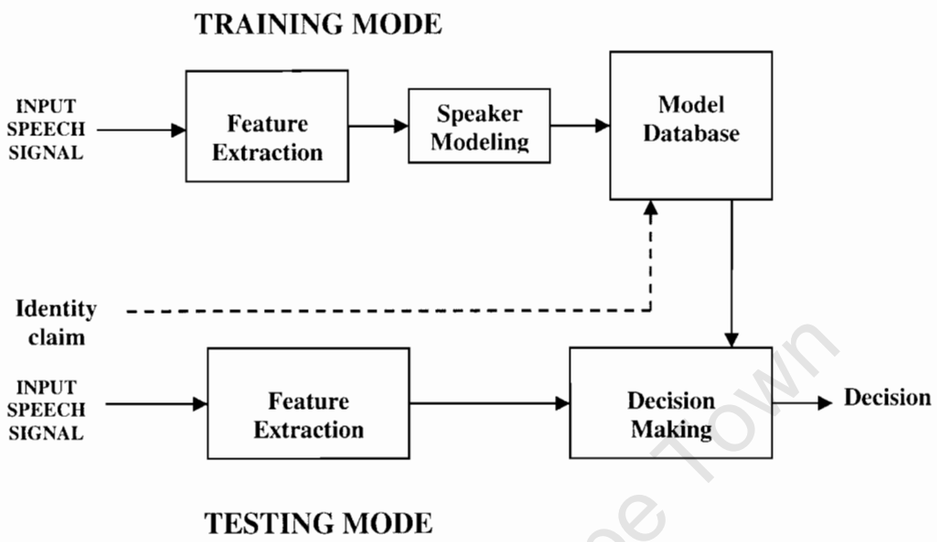

Figure 3-1: A generic speaker recognition system

Figure 3-1 illustrates the two modes of operation as well as the pattern recognition aspects mentioned previously, each represented as an independent component. However, in speaker recognition terminology, data acquisition is referred to as *feature extraction* or *speech parameterization* (which occurs in the *front-end*). Data representation is referred to as *speaker modeling* and decision-making is often referred to as *classification* (which together with speaker modeling occurs in the *back-end*). In the training mode, new speakers are enrolled into the system and in the testing mode the recognition of the speakers takes place [26].

The purpose of feature extraction is to convert a raw speech signal into a compact and efficient representation that is more stable and discriminative than that of the original signal. The output of this component is a sequence of feature vectors where the

individual elements of each feature vector are known as *features*. Feature extraction takes place in both the training and testing modes. In this research *Mel-frequency Cepstral Coefficients* (MFCC) are produced by the feature extraction component. These features are aimed at emulating the spectral compression applied by the human auditory system to an incoming speech signal [28] and, are the most commonly used features in speech-related research. For speech recognition for example, much of what one does for features is in the eradication of clues as to who is speaking thereby enabling the design of speaker independent systems. However, MFCCs contain enough speaker specific information for use in speaker recognition systems. In particular, MFCCs have been shown to be very effective for text-independent speaker identification [29]. In the training mode the features generated from the input speech signal are directed into the speaker modeling component where models for each speaker's speech characteristics are created. This component uses a *Gaussian Mixture Model* (GMM) to represent each speaker's speech characteristics. GMMs are a statistical modeling technique that can be used to represent the underlying distributions of the MFCCs, generated by the feature extraction component, for each speaker, as they have the ability to model arbitrary densities. For any speaker, his (or her) GMM is a weighted linear combination of M component unimodal Gaussian densities $b_i(\mathbf{x})$ each parameterized by a mean vector $\boldsymbol{\mu}_i$ and covariance matrix Σ_i. These parameters are collectively represented as $\lambda = \{p_i, \boldsymbol{\mu}_i, \Sigma_i\}$ $i = 1, ..., M$ where p_i are the mixture weights and satisfy the constraint $\sum_{i=1}^{M} p_i = 1$. A GMM computation is given by the following equation

$$P(\mathbf{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\mathbf{x}) \tag{3.1}$$

where

$$b_i(\mathbf{x}) = (1/(2\pi)^{D/2} \mid \Sigma_i \mid^{1/2}) \quad \mathrm{x} \quad \exp\left(\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)' \Sigma_i (\mathbf{x} - \boldsymbol{\mu}_i)\right) \tag{3.2}$$

and $\mathbf{x}$ is a D-dimensional feature vector [30].

In so doing, the system "learns" the speaker's voice. The models generated by the speaker modeling component are subsequently stored in a model database for use

during testing. During testing, the decision-making component compares the features generated by the feature extraction component to the speaker models stored in the model database and a measure of similarity (usually a numerical value) is computed. Depending on the task, the decision making component uses these values to either assign a speaker identity to the speech signal or to verify that it belongs to a particular speaker [26].

3.1.1 Speaker identification

Given a sample of speech, speaker identification is the task of deciding who, among a finite set of enrolled speakers, produced it [31]. The test utterance is scored against all possible speaker models, and the model that produces the highest score determines the speaker's identity. The task involves making a 1: N classification, where N is the number of enrolled speakers. One of the main limitations of speaker identification is that as N increases, the probability of correctly identifying a speaker decreases [29]. In addition to the decrease in accuracy, the size of N also adversely affects the execution time of the system, that is, the larger the number of enrolled speakers, the longer the execution time [26].

3.1.2 Speaker verification

Given a sample of speech and an identity claim, speaker verification (also known as speaker detection or speaker authentication) is the task of determining whether the sample of speech can be attributed to the enrolled speaker associated with the identity claim or not [31, 32]. This is done by testing the model of the targeted speaker with the utterance, comparing the score obtained to a threshold, and deciding on the basis of this comparison whether or not to accept the claimant. Therefore, a speaker verification system has to make a binary decision as to whether the identity is accepted or rejected. When accepted, the claimant is referred to as either the *legitimate* or *target speaker* and when rejected either an *imposter* or *non-target speaker* [31]. Unlike speaker identification, speaker verification performance is not dependent on the number of potential imposters, that is, the number of enrolled speakers. However, the composition of the imposter set will naturally affect performance if imposters similar to the targeted speaker(s) are selected [26].

Speaker recognition systems can also be classified as being either *text-dependent* or *text-independent*. The former requires the speaker to utter key words; phrases or sentences having the same textual content for both training and recognition trials, whereas the latter does not rely on specific textual data for training and testing. Most commercial systems in operation today are text-dependent, as the additional knowledge of the specific phrase to expect can be used to enhance security. This is done by simultaneously making use of speech recognition to verify the text of the input phrase [33]. However, text-independent speaker recognition systems are more flexible than text-dependent ones as recognition can be performed in the background, that is, regardless of the spoken utterance and without explicit user co-operation, while users are engaged in other speech interactions [32, 34].

3.2 Speaker recognition with microphone arrays

While much research has been conducted into the use of microphone arrays with *speech recognition* systems (see [1, 3, 5, 6, 10, 35]), very little work has been done for speaker recognition tasks. In 1994, Lin [7] investigated the use of microphone arrays with speaker recognition, using a matched-filter array with a vector quantization based speaker identification system. While their results showed significant performance improvements in noisy conditions, the research is at least partially outdated by the recent shift to *Gaussian mixture model* (*GMM*) speaker recognition systems [18]. More recently, a number of research papers investigating the use of microphone arrays in GMM based speaker recognition systems in noisy conditions have been produced. Most of the research demonstrates the benefits of using a microphone array over a single microphone in hands-free speaker recognition (e.g. [36]).

One of the limitations of current research is the lack of results for speaker verification. As we well know, speaker recognition applications can be categorized as either identification or verification tasks. To date, most of the research in microphone array speaker recognition has been confined to the task of speaker identification. Therefore, while some research has been conducted in the field of speaker recognition using microphone arrays, this has been minimal, and to further research in the field a number of issues should be addressed, including:

1. Investigating the use of more sophisticated beamforming techniques.
2. Looking at more realistic methods of generating multi-channel speech databases for experiments, that is, actual microphone array recordings.
3. More research into the use of microphone array enhanced speech with state-of-the-art GMM based speaker recognition systems.
4. Conducting experiments into the effect of microphone array enhanced speech on speaker verification performance.

The experimental work done in this study aims at addressing the above issues, firstly by proposing a beamforming technique that targets diffuse noise in a real office environment. Secondly, although speech from the TIMIT database is used, it is played back and recorded using an actual four element microphone array to create a realistic database. Also, the microphone array is evaluated on a GMM based speaker recognition system that uses *Mel-frequency cepstral coefficients* (MFCC) that have been shown to perform reasonably well on numerous speaker recognition tasks [31, 37, 38]. Lastly, the experimental evaluation in chapter 6 looks at both speaker identification and speaker verification performance.

3.3 Summary

This chapter provided a general introduction and overview of speaker recognition systems, illustrating the key modules that make up such systems. Speaker identification and speaker verification systems were differentiated and briefly discussed. Some of the limitations that this thesis hopes to address regarding the effective implementation of microphone arrays with speaker recognition systems were looked at. The following chapter introduces noise canceling and its implementation in beamforming. It also discusses the design and implementation of a post-filter to further improve the system.

Chapter 4

4 Noise canceling beamformer with wiener post-filter

This chapter describes the Generalized Sidelobe Canceller (GSC) noise canceling beamformer with a Wiener post-filter. The technique was proposed for this study to improve microphone array beamforming for speaker recognition applications by decreasing the level of noise and the amount of signal distortion in both training and testing data.

4.1 Motivation for using Noise canceling

With speech processing techniques being increasingly applied in real noise environments, speech enhancement is currently an important area of research. Several approaches exist, each taking into account different types of knowledge about the desired signal, such as speech production models or spectral content [39]. This thesis focuses on the use of spatial information to enhance the desired signal by using a microphone array. Microphone array beamforming techniques allow spatial filtering of a noisy input, enhancing speech from a desired direction while attenuating noise from all other directions.

There are some good reasons for using the standard delay-and-sum beamformer for speech input tasks; e.g. where artifacts added by adaptive techniques are detrimental. However, it is well known that this beamformer is not well suited to the task of

speech enhancement for some applications, and that many other techniques have been proposed. Among these, the Generalized Sidelobe Canceller (GSC), discussed in chapter 2, suggested by Griffiths and Jim [21] has shown promising performance in noise reduction for practically sized microphone arrays while maintaining low speech signal distortion. However, in a diffused noise field the noise reduction is less significant and the performance is further degraded when the noise signal is non-stationary.

In this contribution the thesis seeks to build upon the standard generalized sidelobe canceller by modifying the noise canceling path in the system and thus presenting a *superdirective* beamformer for higher directivity and increased performance in a diffuse noise field. The adaptive noise canceller enables the array to adapt to varying noise conditions, providing attenuation to undesired noise sources and leading to lower noise power in the beamformed output. The usual method of estimating a signal corrupted by additive noise is to pass the composite signal through a filter that tends to suppress the noise while leaving the desired signal relatively unchanged. The design of such processing techniques is the domain of optimal filtering, which originated with the pioneering work of Wiener and was extended and enhanced by the work of Kalman, Bucy and others [40-44].

Noise canceling is a variation of optimal filtering that is highly advantageous in many applications. It uses an auxiliary or reference input derived from one or more sensors located at points in the noise field where the desired signal is weak or, in the case of microphone arrays, from the output of a blocking matrix from which the desired signal has been blocked. This input is filtered and subtracted from the primary input which contains both the desired signal and noise. As a result, the primary noise is attenuated or eliminated by cancellation. In circumstances such as microphone arrays, where noise canceling is applicable, a degree of noise rejection can be achieved that would be difficult to achieve by direct beamforming or filtering [45]. The basic noise canceling situation is illustrated in Figure 4-1. A signal is transmitted to a sensor that receives the signal plus an uncorrelated noise, n_0. The combined signal and noise, $s + n_0$, form the primary input to the canceller. A second sensor receives a noise n_1, which is uncorrelated with the signal but correlated in some unknown way with the noise n_0. This sensor provides the reference input to the canceller. The noise n_1 is filtered to

produce an output y that is a close replica of n_0. This output is subtracted from the primary input $s + n_0$ to produce the system output, $s + n_0 - y$.

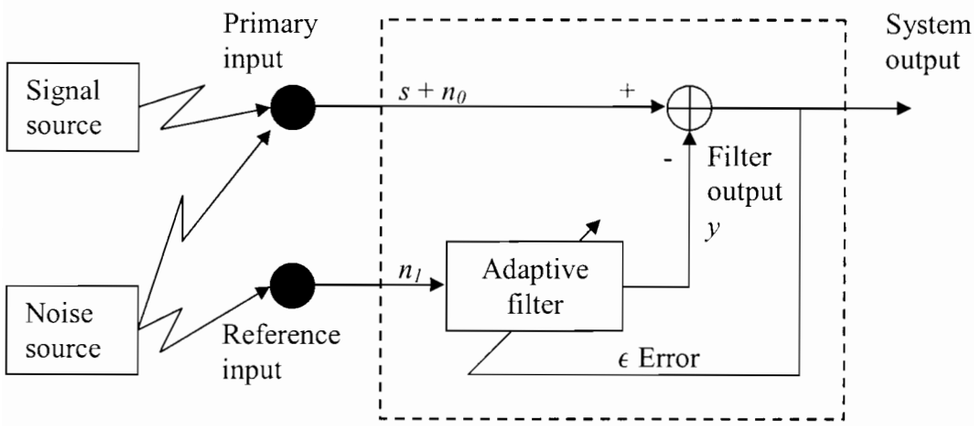

Figure 4-1: Adaptive noise canceling concept [45]

The use of a suitable single channel post-filter to further enhance the desired signal is also investigated. A thorough study of the use of post-filters with microphone arrays has been undertaken by a number of researchers (see [22, 39]) who have shown how such a post-filter can enhance the performance of a beamformer.

4.2 Beamformer design

As mentioned in the previous section, the GSC noise canceling beamformer presents some superdirectivity to the system. The term directivity describes the ability of a beamformer to suppress noise coming from all directions without affecting a desired signal from one principal direction [4]. The GSC (Figure 2-15) includes a blocking matrix that removes signals arriving from the look direction to produce reference signals free of the desired signal. The reference signals are input to adaptive filters performing unconstrained noise minimization. The adaptive filter outputs are then subtracted from a delayed primary signal consisting of the desired signal plus interference. The unconstrained minimization of the noise can be achieved in the time-domain using the least mean squares (LMS) algorithm, which is favoured

because of its low computational complexity. Because adjustment of the adaptive weights is proportional to the desired signal strength even in an ideal situation, and because some of the desired signal leaks into the reference noise signal under realistic conditions (due to sensor mismatch, mis-steering, or reverberation), the traditional GSC performs poorly, degrading the desired signal.

A number of modifications and adjustments to the GSC effectively overcome these problems. One of these modifications involves preventing the cancellation of the desired signal based on its reflections by appropriate selection of the primary signal delay [46]. In particular, the delay must be shorter than the interval between the arrival of the direct wave and the first reflection at the microphones [47]. Another modification would be to modify the noise canceling path of the beamformer so it can adapt to diffuse noise or measured noise fields of particular environments.

In order to design an optimal beamformer, we have to minimize the power of the output signal *y(t)* of the array. The output *power spectral density* (PSD) is given by

$$\Phi_{YY} = \boldsymbol{W}^{H} \boldsymbol{\Phi}_{XX} \boldsymbol{W}, \tag{4.1}$$

where

$$\boldsymbol{\Phi}_{XX} = \begin{pmatrix} \Phi_{X_0X_0} & \Phi_{X_0X_1} & \cdots & \Phi_{X_0X_{N-1}} \\ \Phi_{X_1X_0} & \Phi_{X_1X_1} & \cdots & \Phi_{X_1X_{N-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{X_{N-1}X_0} & \Phi_{X_{N-1}X_1} & \cdots & \Phi_{X_{N-1}X_{N-1}} \end{pmatrix} \tag{4.2}$$

is a power spectral density matrix of the array input signals. $\boldsymbol{W}$ represents the matrix of the frequency-domain coefficients of the beamformer sensors and $\boldsymbol{W}^{H}$ its conjugated transposition (Hermitian operator). The PSD is a function of the input signal and the coefficients we want to determine. In order to avoid the trivial solution $\boldsymbol{W} = 0$, the minimization is constrained to give an undistorted signal response in the desired look direction, i.e.,

$$\boldsymbol{W}^{H}\boldsymbol{d} = 1 \tag{4.3}$$

where $\boldsymbol{d}$ is the representation of the delays and the attenuation in the frequency domain, which depends on the actual geometry of the array and the direction of the signal source. Therefore, the following minimization problem has to be solved:

$$\min_{\boldsymbol{W}} \boldsymbol{W}^{H}\boldsymbol{\Phi}_{XX}\boldsymbol{W} \quad \text{subject to } \boldsymbol{W}^{H}\boldsymbol{d} = 1 \tag{4.4}$$

Since we are only interested in the optimal suppression of the noise and we assume a perfect correspondence between the direction of the desired signal and the look direction of the array, only the noise PSD matrix $\boldsymbol{\Phi}_{NN}$ is used. The well known solution for Equation (4.4) is called the *Minimum Variance Distortionless Response* (MVDR) beamformer [48] and is given by

$$\boldsymbol{W} = \frac{\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{d}}{\boldsymbol{d}^{H}\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{d}} \tag{4.5}$$

and can be derived using the Lagrange multiplier [49] or gradient computation [50]. Assuming a homogenous noise field the solution is a function of the coherence matrix

$$\boldsymbol{W} = \frac{\boldsymbol{\Gamma}_{NN}^{-1}\boldsymbol{d}}{\boldsymbol{d}^{H}\boldsymbol{\Gamma}_{NN}^{-1}\boldsymbol{d}} \tag{4.6}$$

The coherence matrix given by

$$\boldsymbol{\Gamma}_{NN} = \begin{pmatrix} 1 & \Gamma_{N_0N_1} & \cdots & \Gamma_{N_0N_{N-1}} \\ \Gamma_{N_1N_0} & 1 & \cdots & \Gamma_{N_1N_{N-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{N_{N-1}N_0} & \Gamma_{N_{N-1}N_1} & \Gamma_{NN} & 1 \end{pmatrix} \tag{4.7}$$

allows an easier examination of beamformers for different noise fields, since many theoretically defined noise fields can be expressed by their coherence functions [4]. Equation (4.5) or (4.6) can be interpreted as a spatial decorrelation process followed

by a matched filter for the desired signal. The normalization in the denominator leads to unity signal response for the look direction.

The design procedure reduces to the choice of theoretically well defined noise fields in order to get optimal designs for different applications. Furthermore, different models for the desired signal can be included, leading to far-field and near-field designs. The desired signal model for the standard far-field design for linear arrays with equidistant sensors is given by

$$\begin{aligned} \boldsymbol{d}^T = [1, &\exp(-j\Omega f_s c^{-1} l\cos(\theta_0)), \exp(-j\Omega f_s c^{-1} 2l\cos(\theta_0)), \\ &\cdots, \exp(-j\Omega f_s c^{-1}(N-1)l\cos(\theta_0))] \end{aligned} \tag{4.8}$$

where l is the inter-sensor spacing, Ω, the wavefront frequency and f_s denotes the sampling frequency.

4.2.1 Delay-and-Sum beamformer

The well known delay-and-sum beamformer is included for comparison purposes. It is an optimal beamformer for optimizing the *white noise gain* (WNG), which is spatially uncorrelated noise that can be caused by self-noise of the sensors [4]. The coefficients are derived from Equation (4.6) by inserting the coherence matrix for spatial uncorrelated noise $\boldsymbol{\Gamma} = \boldsymbol{I}$. Thus,

$$\boldsymbol{W} = \frac{1}{N}\boldsymbol{d} \tag{4.9}$$

which is similar to Equation (2.57) in the time domain. The white noise gain is optimal in this case and reaches N, the number of sensors.

In order to optimize the directivity factor, which depends on the diffuse noise field, Equation (4.6) has to be solved by using the coherence matrix of the diffuse noise field, given by

$$\Gamma_{NN}(e^{j\Omega})\Big|_{Diffuse} = \frac{\sin(\Omega f_s l/c)}{\Omega f_s l/c} = \sin c\left\{\frac{\Omega f_s l}{c}\right\} \tag{4.10}$$

where Ω represents the frequency of a single wavefront, f_s denotes the sampling frequency, c is the speed of sound and l the distance between the array sensors. The resulting coefficients represent the superdirective beamformer. Figure 4-2 illustrates the beampattern of a delay-and-sum and a superdirective beamformer. The x-axis represents the incoming spatial angle $(0\cdots 2\pi)$ and the y-axis represents the frequency of the signal in kHz. This grey-scaled image represents the attenuation of the incoming signals in dB. The delay-and-sum beamformer is unable to suppress low frequency noise coming from any direction. In contrast, the superdirective beamformer attenuates noise coming from directions other than the look direction over the whole frequency range.

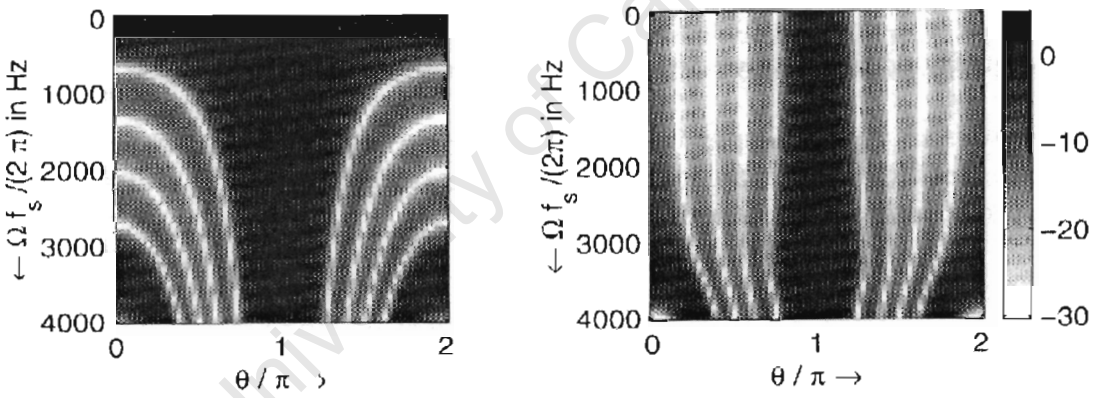

Figure 4-2: Left: beampattern of a delay-and-sum beamformer. Right: beampattern of an optimal array for diffuse noise (superdirective beamformer). (l = 5 cm, N = 5) [4]

4.2.2 Measured noise fields

Thus far, only data-independent designs have been considered. However, if *a priori* information of the noise field is available, it should be used to improve performance. For example, this information could be a prescribed direction (θ = angle) of an incoming noise source. Assuming the noise source is in the far field of the microphone array, the complex coherence function between two sensors is given by

$$Re\left\{\Gamma_{X_i X_j}(\omega)\right\} = \cos\left(\frac{\Omega f_s \cos(\theta) l_{ij}}{c}\right) \tag{4.11}$$

$$Im\left\{\Gamma_{X_i X_j}(\omega)\right\} = -\sin\left(\frac{\Omega f_s \cos(\theta) l_{ij}}{c}\right) \tag{4.12}$$

Inserting the coherence matrix into Equation (4.6) forms a null in the direction of the noise source over the whole frequency range.

In addition, if we assume stationarity, the actual noise field can be measured and we can solve the design equation which results in the minimum variance distortionless solution. Adaptive algorithms like the constrained projection by Cox [48], or the original algorithm by Frost [49], will converge exactly to the same solution under the assumption of stationary noise [4].

4.2.3 GSC implementation

Assuming a time aligned input signal, this would mean that the look direction vector ***d*** would be replaced by the column vector

$$\mathbf{1} = [1,1,\cdots 1]^T$$

containing N ones, and the PSD matrix or the coherence matrix containing the statistical information after time alignment. Figure 4-3 illustrates this.

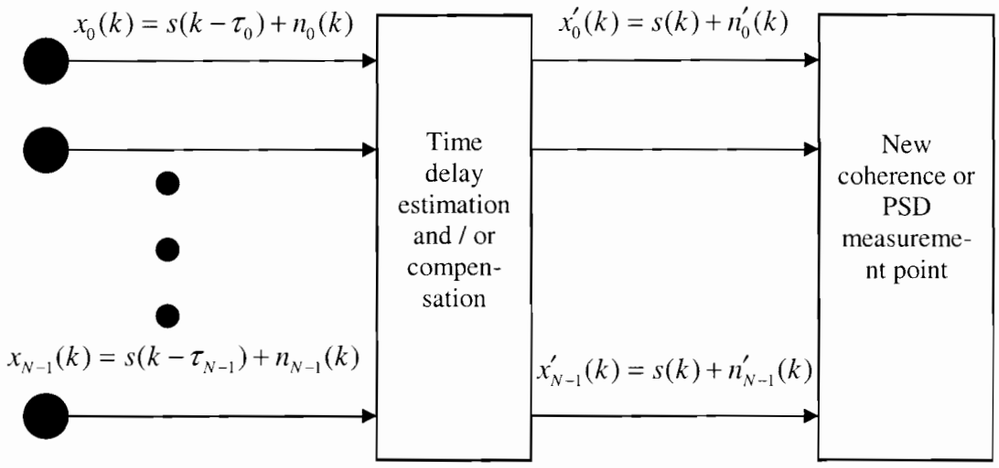

Figure 4-3: Signal model after time delay compensation

This gives the solution

$$\boldsymbol{W}\Big|_{\substack{time\\aligned}} = \frac{\mathbf{1}^T(\boldsymbol{\Gamma}'_{NN} + \mu\boldsymbol{I})^{-1}}{\mathbf{1}^T(\boldsymbol{\Gamma}'_{NN} + \mu\boldsymbol{I})^{-1}\mathbf{1}} \tag{4.13}$$

where μ is a small scalar added to the main diagonal of the normalized PSD or coherence matrix in order to overcome the problem of self-noise amplification [51]. This solution can be decomposed into two orthogonal parts, following the ideas of Griffiths and Jim [21] (see section 2.4.3). One part represents the constraints only and the other part represents the unconstrained coefficients to minimize the output power of the noise. Figure 4-4 illustrates the decomposed structure. The multi-channel time-aligned input signal $\boldsymbol{X}$ is multiplied by $\boldsymbol{W}^C$ (channel filters) to execute the constraints. Additionally, the input signal is directed into the noise-only path where the desired signal is spatially filtered out by a blocking matrix $\boldsymbol{B}$. The resulting vector X^B is processed by the *adaptive linear combiner* (ALC) and then subtracted from the output of the upper path of the structure to get the noise-reduced output signal Z. Some authors [21, 52, 53], have shown the likeness between this structure and the delay-and-sum beamformer if

$$\boldsymbol{W}^{C} = \frac{1}{N}\mathbf{1}. \tag{4.14}$$

In addition, the matrix $\boldsymbol{B}$ has to fulfill the following properties:

1. The size of the matrix is $(N-1)$ x N
2. The sum of all values in one row is zero
3. The matrix has to be of row rank N – 1.

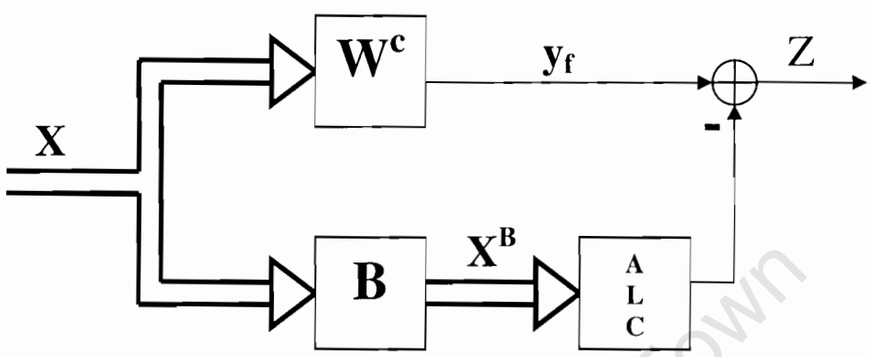

Figure 4-4: Model of the decomposition of the optimal weight vector into two orthogonal parts [4]

An example of a blocking matrix for $N = 4$ is given by

$$\boldsymbol{B} = \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix} \tag{4.15}$$

Also, a well known example of a blocking matrix is the Griffith-Jim matrix which subtracts two adjacent channels only

$$\boldsymbol{B} = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 1 & -1 \end{pmatrix} \tag{4.16}$$

The final step is the computation of the adaptive linear combiner (ALC). The ALC, or nonrecursive adaptive filter is fundamental to adaptive signal processing. Figure 4-5 illustrates the general form of the adaptive linear combiner. There is an input signal matrix with vector elements x_0, x_1, …, x_{N-1}, a corresponding set of adjustable weights, w_0, w_1, … , w_{N-1}, a summing unit, and a single output signal, n. The combiner is called "linear" because for a fixed setting of the weights its output is a linear combination of the input components. From the input signal notation in Figure 4-5, we obtain the input-output relationship as

$$n = \sum_{i=0}^{N-1} w_i x_i \tag{4.17}$$

The ALC uses a "desired response" signal at its output to derive an error signal as shown in Figure 4-6.

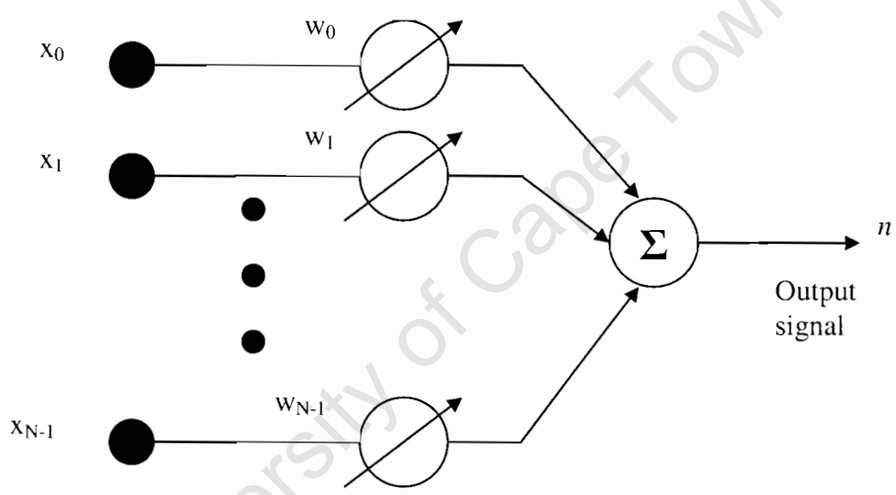

Figure 4-5: General form of adaptive linear combiner

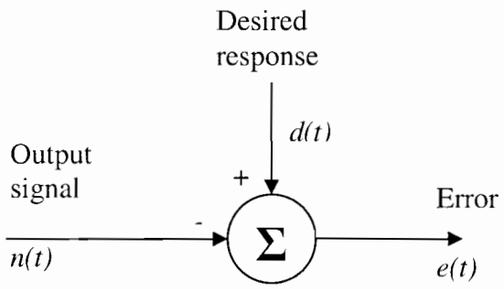

Figure 4-6: Desired response and error signals for ALC

The output signal, *n(t)*, is simply subtracted from the desired signal, *d(t)*, to produce the error signal, *e(t)*.

4.3 Post-filter

In speech enhancement applications, microphone array post-filtering allows for additional reduction of noise components at a beamformer output. To improve the performance of the array, a post-filter is associated with the beamformer. In the context of microphone arrays, the term post-filtering denotes *the post-processing of the array output by a single-channel noise suppression filter* [4]. By incorporating a post-filter with a beamformer, we are able to use the knowledge obtained in spatial filtering to also allow effective *frequency* filtering of the signal. In using both spatial and frequency domain enhancement we are making maximal use of our information about the signal, this being solely its direction of arrival [35]. It is expected that such a combined technique should be capable of yielding results that outperform a system using solely spatial or frequency filtering techniques relying on the same *a priori* signal information.

Figure 4-7 illustrates the fully developed beamforming structure with a post-filter. The post-filter used is based on the commonly used post-filter proposed by Zelinski [54]. The Zelinski post filter uses the input channel auto- and cross-spectral densities to estimate a Wiener post-filter to be applied to the beamformer output. The use of such a post-filter with standard microphone arrays has been thoroughly investigated by Marro *et al* [22], and has been used successfully in a number of speech enhancement and speech recognition applications [23]. The single-channel Wiener

post-filter transfer function proposed for the microphone array implemented in this research work is given as [4]

$$h(f) = \frac{\phi_{ss}(f)}{\phi_{ss}(f) + \phi_{nn}(f)} \tag{4.18}$$

where $\phi_{ss}(f)$ and $\phi_{nn}(f)$ are the auto-spectral densities of the desired speech signal and the undesired noise component respectively.

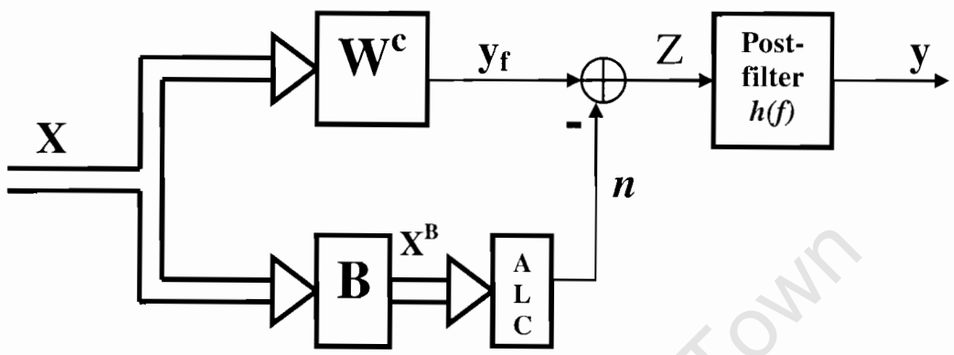

Figure 4-7: Beamformer structure with post-filter

Designing an array beamformer for defined noise fields is dependent on spatial characteristics described by the coherence function. Most of the array characteristics like the beam pattern or the directivity index are directly connected to the coherence function. The design presented in this chapter can be implemented using the general GSC structure as a building block. The structure allows us to combine superdirective beamformers with a post-filter for further noise reduction [55]. The GSC structure on its own shows impressive noise reduction abilities in directional and measured noise fields. The addition of a post-filter provides the necessary enhancement for diffuse noise fields while maintaining low speech distortion.

4.4 Summary

This chapter demonstrated how noise canceling can effectively be implemented using microphone arrays and how variations of the generalized sidelobe canceller can be

used to achieve signal enhancement in different noise fields. By determining the noise field in a particular environment, the design and parameters of the adaptive noise canceling filters can be varied to optimize the beamforming process. This chapter also presented a review of the single channel Wiener post-filter. The Wiener post-filter has been used extensively in speech and microphone array applications, and has been known to improve the robustness of numerous microphone array systems. The following chapter describes the design of the microphone array experimental framework.

Chapter 5

5 Microphone Array Design and Setup

The whole microphone array processing system can be broken down into a number of smaller units. These units are shown in Figure 5-1. The key components of each unit are described in detail in the sections that follow. The whole system basically gathers audio signals through the microphones and amplifies the input using audio amplifiers. The signals are then captured by a computer via a data acquisition card, which digitizes the signals. Beamforming is then performed using MATLAB and the resulting output is either played back as an audio signal or used as input to a speaker recognition system. This chapter describes the layout of the array processing system using a computer to collect data from a 4-sensor microphone array.

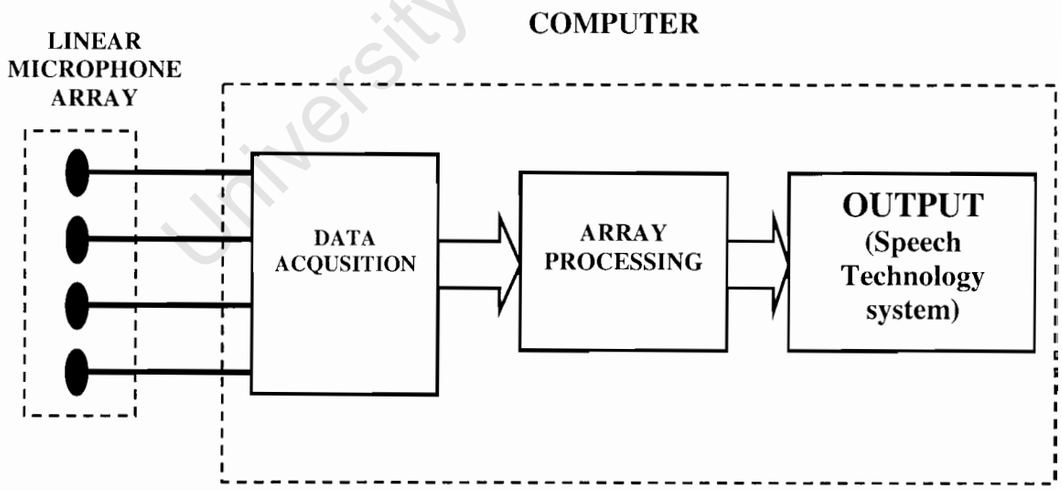

Figure 5-1: Microphone array system

5.1 Hardware Design

The purpose of the hardware developed for this project is to acquire acoustic samples from four microphones simultaneously. This section details the major hardware components used in the data acquisition unit and their layout.

5.1.1 Microphones

A microphone is a transducer that converts acoustic energy to electrical energy. There are five key types of commonly used microphones. These are

1. Moving coil
2. Ribbon
3. Condenser
4. Electret
5. Crystal

All five microphones employ different mechanisms for converting sound energy to electrical energy. As a result all of them have different advantages and disadvantages, which implies that the right type of microphone needs to be selected for the right type of application.

When choosing a microphone for this research, the size, pressure sensitivity and frequency response are some of the parameters that were considered. A microphone's pressure sensitivity is defined as the voltage generated in response to a certain pressure input and is denoted as

$$M_0(\text{Volts/Pascal})$$

The frequency response of a microphone is the characteristic graph obtained by recording the voltage output level in dB while the microphone is exposed to a range of sinusoidal tones of equal intensity. The frequency response is often given as a graph or stated as a variation within a given range of frequency[56].

In addition, there are two different styles of microphones, *unidirectional* and *omnidirectional*. Both receive vibrations from outside sources and convert them to electrical energy. However, unidirectional microphones only pick up sounds aimed directly into their centers. This is ideal for one speaker, but is not as useful when sounds are captured from different directions or distant sources.

On the other hand, omnidirectional microphones can pick up sounds virtually from any direction. However, omnidirectional microphones do have some drawbacks. Due to the fact that omnidirectional microphones cannot discriminate between wanted and unwanted sounds, ambient noise from the environment can be picked up and amplified [57].

The microphones used in this research are omnidirectional *electret* microphones with a flat frequency response range of 20 Hz to 20 kHz and pressure sensitivity of 20mV/Pa at 1 kHz. Their small size is an advantage, allowing arrays of multiple microphones to be built while minimizing the overall size of the array.

Each of the four microphones has a built in charge and requires a few volts of DC power to power the built-in FET in each microphone. Incoming sound pressure changes the separation distance between the sides of the capacitor which in turn changes the capacitance and the output of each microphone. A typical electret condenser microphone capsule is a 2 terminal device (there are also 3 pin capsules) which approximates to a current source when biased with around 1-9 volt and routinely consumes less than half a milliamp. This power is consumed by a very small preamplifier built into the microphone capsule which makes the conversion of very high impedance source of the electret element itself and the cable which needs to be driven [58]. Two separate power supply units were used to provide power to the four microphones, two microphones connected to each power supply unit. This was done to maintain maximum isolation for the purpose of ensuring low crosstalk which is especially important for beamforming microphones as excessive crosstalk might cancel out the beamforming effect.

5.1.2 Amplifiers

Amplification of the incoming audio signals is necessary, firstly because the signals needed to be increased to the input level required by the data acquisition card. Secondly, each microphone, although of the same model, had different characteristics. It was therefore necessary to adjust all of the microphone signals to the same level by adjusting and normalizing the gain of each signal using audio amplifiers.

Thus, the output of each of the four microphones is sent to a LM386N-3 audio amplifier. Figure 5-2 illustrates the connectivity of two of the microphones in the array showing both the power supply and amplifier wiring.

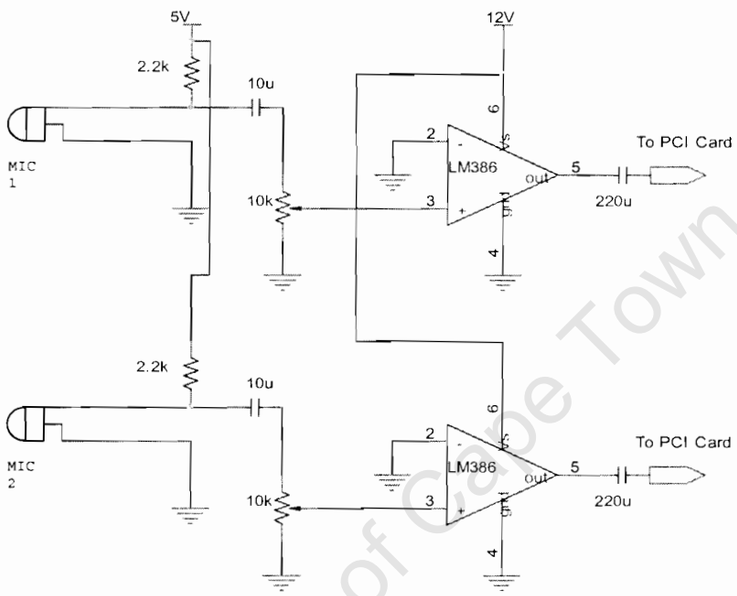

Figure 5-2: Wiring diagram for two of the microphones in the array

The amplifiers used in this design were the LM386N-3 *Low Voltage Audio Power Amplifiers*. The LM386N-3 is a power amplifier designed for use in low voltage consumer applications. The gain is internally set to 20 to keep external part count low and the inputs are ground referenced while the output is automatically biased to half the supply voltage. Other features include a wide supply voltage range and low distortion [59], which is one of the influencing factors in selecting this amplifier. Figure 5-3 illustrates the low signal distortion over the frequency range; 20Hz to 20 kHz.

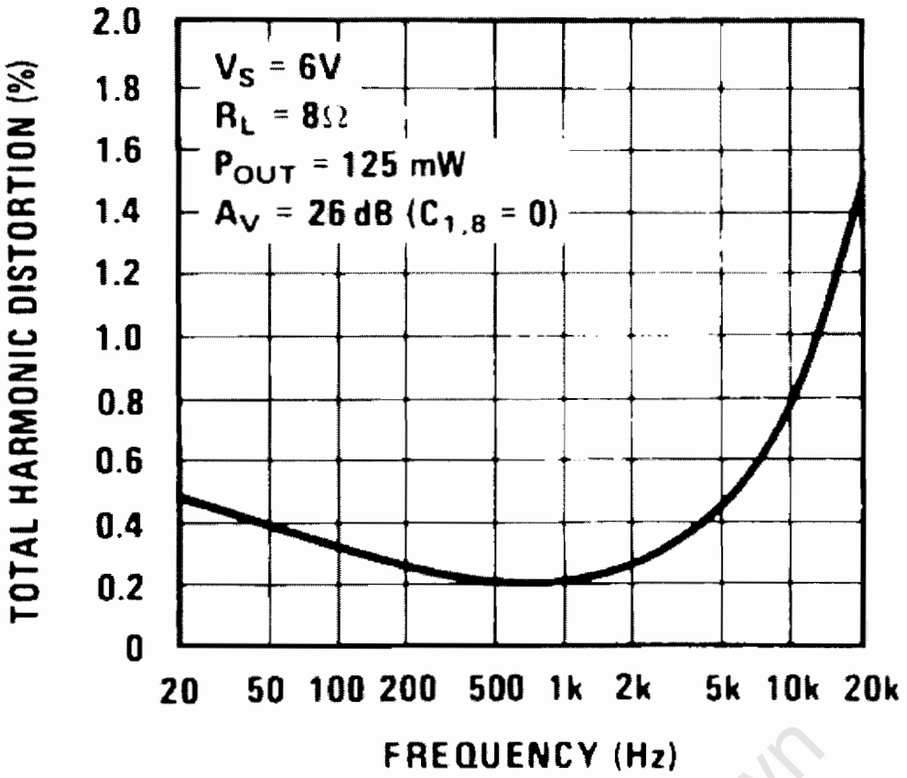

Figure 5-3: Distortion vs. Frequency curve for the LM386 amplifier [59]

5.1.3 Data acquisition

After the amplification stage, the signals are captured by a computer via an analog and digital input/output board in form of a PCI card. The data acquisition board used in this research is the PCI-703-16 manufactured by Eagle Technology. The PCI-703-XX series are 32-bit PCI bus architecture data acquisition boards. They are available in four models, that is the 16, 32 and 64 channels analog input boards as well as the sample-and-hold version. The PCI-703-XX contains digital input and output ports, onboard counters, a frequency generator, analog-in and analog-out sub-systems. It is a multi-purpose analog board that can be used in many applications and is recommended for its low noise and high performance [60].

The PCI-703-16 is a 14-bit, 400 kHz analog input board for PCI based systems. The board features 16 differential analog inputs, 18 digital I/O lines and 3 counters fitted as standard. Included with the board is *WaveView for Windows Data Acquisition and Logging* software, which was the data acquisition software used in this research. All

operating system drivers, utility and test software were also supplied by Eagle Technology.

Once the hardware is set up, WaveView is used to control the data acquisition process and to make the data samples available to programs written in MATLAB.

5.2 Software

5.2.1 Software design

Array processing uses a set of inputs arranged in a known spatial pattern. A linear array is used in this work as it simplifies mathematical calculations that become more involving with increased complexity in the array configuration. Waves propagated towards an array of microphones are spherical in nature. However, if the waves are propagated from a source at a sufficient distance (far field sources), it can be assumed that the waves are nearly planar upon arrival at the microphone array. Let us consider two sound sources, X and Y, that emit sound waves, *x(t)* and *y(t)* respectively, directed towards a four element linear microphone array. Depending on the angle of arrival (θ_x or θ_y), with respect to the array, and the separation distance *d*, of the microphones, there will be a time delay of either t_x or t_y, when each microphone receives the inputs *x(t)* and *y(t)* respectively. The delay for adjacent elements in relation to the angle of arrival can be calculated using the formula

$$\tau_x = \frac{d \cos\theta_x}{c} \tag{5.1}$$

where *c* represents the speed with which the waves propagate and *d* the inter-element spacing. A similar equation holds for source Y. If there are N microphones in the array all equidistantly spaced, then there are N different outputs.

Knowledge of these delays allows the microphone array to focus its beam towards a single source. This is done by applying delays to each microphone so that each output is in phase with a single source whilst out of phase with the other. In order to focus on source X, the delays are applied to each microphone so as to steer the main beam

towards the look angle, θ_x. The outputs are then summed and normalized to yield a single beamformed signal focused on source X. The out of phase summation of the delayed version of *y(t)* from source Y can now be considered to be the noise source *n(t)*. The final result is a signal with improved signal-to-noise ratio and reduced distortion.

This provides the basis for the software requirements for the beamforming algorithms implemented in this research. The hardware discussed in section 5.1.3 works in unison with data acquisition software to acquire the necessary data from the microphone array for processing using the software design principles discussed above. WaveView for Windows is the data acquisition software that acquires and provides the required data for processing.

5.2.2 WaveView

WaveView for Windows is a Microsoft® Windows based Data Acquisition software package which enables the user to collect and analyze data. It is developed by Eagle Technology and supports the PCI-703-16 board. Possible applications for WaveView for Windows include:

1. 1 – 64 channel digital storage scope
2. Strip chart data logger
3. High speed streaming (up to 64 channels)
4. Continuous process monitoring

Data can be streamed to disk at full 400kHz throughput of the PCI-703-16 on a reasonably fast machine and disk [61].

WaveView basically supports two main modes of operation, chart recording and scope mode. The chart recorder was designed for sampling and saving data over long periods of time. It can record data at a rate of a sample per second or as slow as a sample every 10 hours. The chart recorder can record a wide range of data inputs, both analog and digital [61]. When running in the scope mode, data can be viewed on screen in real-time on a voltage versus time axis. The PCI-703-16 is supported by the scope mode, with a sampling rate of up to 400 kHz for analog inputs. While sampling,

data can also be streamed to disk for later use. The scope mode also has the option of exporting data or graphs to popular standards such as a bitmap image or CSV data file. WaveView provides data in formats compatible with a variety of sophisticated display and analysis packages, including Excel and MATLAB [62]. For this work the software was configured to sample data at a rate of 64 kHz (16 kHz per microphone).

5.2.3 MATLAB™

All of the beamforming algorithms were implemented using MATLAB software. Data collected using the data acquisition board and via WaveView is used by the beamforming algorithms implemented in MATLAB to produce enhanced output signals.

MATLAB receives data from WaveView in form of a matrix with four columns, each column representing data from each of the four microphones. The beamforming algorithms have information relating to the direction of the desired source in relation to the array, the spacing between adjacent microphones, the number of microphones and the sampling frequency. The noise canceling beamformer is based on the generalized sidelobe canceller beamformer algorithm. There are three primary components used in the design: a steering delay; a primary signal estimator; and an adaptive filter. The code takes in the signal matrix as its input and upsamples the signals by an appropriate factor, calculates the appropriate delay for each signal and applies the delays. The appropriate weighting factors are also calculated and applied to each signal for the purpose of steering the microphone array towards the look angle. Following the calculation of the delays and the steering, the signals are down-sampled to the original sampling rate. The output of the steering element is a set of signals defined as

$$m_i(t) = s(t) + n_i(t), \qquad i = 1, 2, \ldots, N \tag{5.2}$$

with each signal corresponding to a different receiver, i. In each signal, $m(t)$, the signal at the look angle, $s(t)$, is in phase and all other signals from other directions are out of phase. The adaptive filter estimates the noise signal that comes from the primary signal estimator. However, the filter requires a correlated noise source, $n(t)$. Having N outputs, $N - 1$ correlated sources can be generated by subtracting adjacent

pairs of outputs from the steering element or by using a blocking matrix. Each output from the blocking matrix is then fed into the adaptive linear combiner (ALC) and then summed at the output to estimate the noise.

The adaptive filter attempts to estimate noise at each instance and subtract it from the output by using the correlated noise source *n(t),* and constantly adjusts to minimize the mean square error of the output.

The ALC is a basic unit used in many adaptive systems. It is a tapped delay line for a single input, consisting of a series of delays and a set of corresponding weights. These weights are varied so *n(t)* can estimate the noise sufficiently, and minimize the mean square error [63].

Finally the adaptive filter output is subtracted from the output of the primary signal estimator to give the final output of the whole beamforming process.

5.3 Summary

This chapter describes the design and implementation of the microphone array system evaluated in this research. The system facilitates the acquisition of speech signals for the speaker recognition systems and is used in evaluating the different beamforming algorithms. The following chapter is aimed at experimentally evaluating the noise canceling beamformer using the experimental framework described in this chapter.

Chapter 6

6 Experimental Results

This chapter is intended at experimentally evaluating the noise canceling beamformer described earlier in chapter 4. The results are analyzed and possible explanations are provided. A total of 200 speech samples, comprising of 100 training and 100 testing speech utterances, from the first 100 speakers from the TIMIT Acoustic-Phonetic Continuous Speech Corpus were used. Each utterance was acquired from distances of 50cm and 100cm from the array and perpendicular to it. The speech was recorded in an office environment with interfering noise mainly from air conditioners and other randomly distributed speakers within and around the office. No additional noise was artificially introduced to the data. In section 6.1, objective quality experiments which rely on mathematically based measure between the original and beamformed speech signals are performed. Section 6.2 evaluates the beamformed speech on speaker identification and speaker verification tasks. This is done in order to determine whether or not the proposed noise canceling beamformer has the ability to outperform a single microphone in a non-close-talking, diffuse noise environment for speaker recognition tasks.

6.1 Objective Quality Assessment

This section compares the signal quality of the noise canceling beamformer, the delay-and-sum (DS) beamformer, the generalized sidelobe canceller (GSC) beamformer and single microphone speech to clean speech acquired using a close-talking microphone. The objective quality measure results are presented in two areas; Itakura-Saito (IS) distortion measure and segmental signal-to-noise ratio (SegSNR).

There are several ways of obtaining overall quality scores. For most measures, finding a mean across a large test set is reasonable [64]. Table 6-1 summarizes the results for the two objective measures considered in this research for four beamforming algorithms and single microphone speech.

Table 6-1: Summary of IS and SegSNR measures for single microphone and four beamforming techniques (average ± standard error of mean)

BEAMFORMER	IS	SegSNR (dB)
Single Mic.	7.79 ± 0.88	-7.92 ± 0.04
Delay-Sum	6.99 ± 0.56	-7.92 ± 0.03
GSC	7.69 ± 0.64	-8.00 ± 0.03
NC-GSC	5.36 ± 0.31	-7.49 ± 0.03
NC-GSC+Wiener	5.24 ± 0.20	-6.83 ± 0.06

For both measures, values closer to 0.00 reflect an improvement in quality. It can be seen that all the beamforming algorithms provided some quality improvement compared to single microphone speech, with the noise canceling beamformer with post-filter (NC-GSC+Wiener) outperforming the other three beamformers. The NC-GSC+Wiener beamformer reduces noise at the beamformer stage as well as at the post-filter and minimizes distortions to the signal. The effect of the post-filter is seen when the distortion and segmental SNR results of the NC-GSC and the NC-GSC+Wiener beamformers are compared.

6.1.1 Distortion measure

For an original clean frame of speech with *linear prediction* (LP) coefficient vector, $\overrightarrow{a_\phi}$, and processed speech coefficient vector, $\overrightarrow{a_d}$, the Itakura-Saito distortion measure is given by,

$$d_{IS}(\overrightarrow{a_d},\overrightarrow{a_\phi}) = \left[\frac{\sigma_\phi^2}{\sigma_d^2}\right]\left[\frac{\overrightarrow{a_d} R_\phi \overrightarrow{a_d^T}}{\overrightarrow{a_\phi} R_\phi \overrightarrow{a_\phi^T}}\right] + \log\left(\frac{\sigma_d^2}{\sigma_\phi^2}\right) - 1 \qquad (6.1)$$

where σ_d^2 and σ_ϕ^2 represent the all-pole gains for the processed and clean speech frame respectively and R_ϕ represents the input autocorrelation matrix [64].

When noise reduction is considered, we normally think of improving the signal-to-noise ratio (SNR). This may not be the most appropriate performance criterion for speech enhancement. All listeners have an intuitive understanding of speech quality, intelligibility and listener fatigue [64]. The Itakura-Saito distortion measure shows the level of distortion for each frame across time. Figure 6-1 shows a plot of average Itakura-Saito distortion measure for utterances from 10 randomly chosen speakers from the first 100 speakers of the TIMIT database. The distortion measure is a comparison between beamformed speech and clean speech. Values close to zero show low distortion as this is basically a difference in distance measure between clean speech and enhanced speech. Figure 6-1 shows that the NC-GSC+Wiener beamformer is more stable as can be seen from the minimal variance of the results, and it generally outperforms the other techniques in this regard.

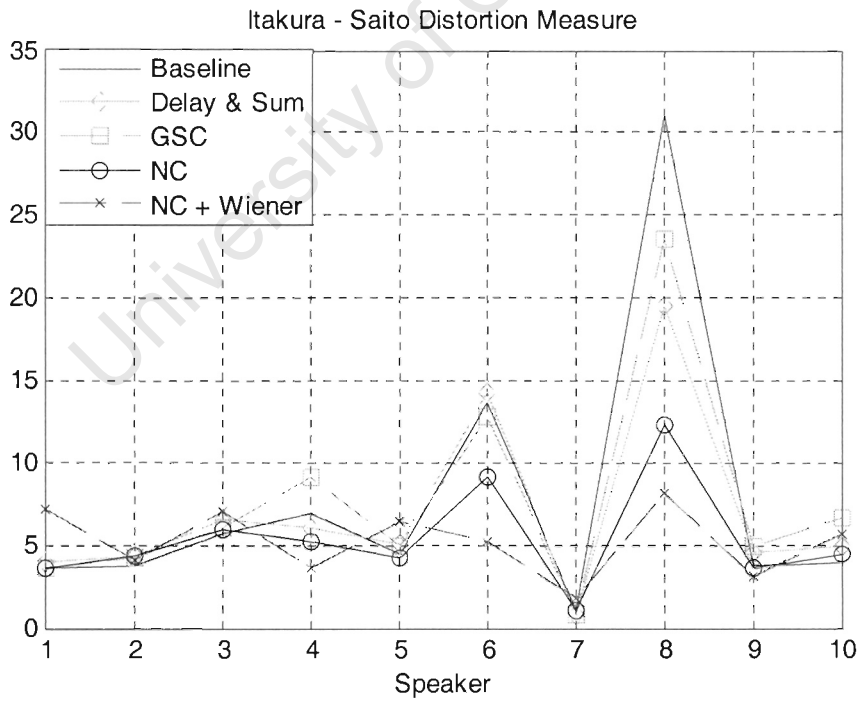

Figure 6-1: Overall mean IS quality measure for 10 speakers

Another way to compare performance is by using quality measure histograms as shown in Figure 6-2. In the histograms the x-axis represents the extent of the

deviation of the processed signal from the original signal (distance of processed speech from clean speech) and the y-axis represents a count of the number of times each valued deviation occurs. For the Itakura-Saito measure distribution, it is seen that after beamforming, the algorithm moves the degraded frames closer to the noise free '0' distortion (towards zero on the x-axis). The important aspect here is to compare the number of frame tails of the distributions and the number of occurrences near the zero distortion, thus reflecting the consistency of the beamforming algorithm.

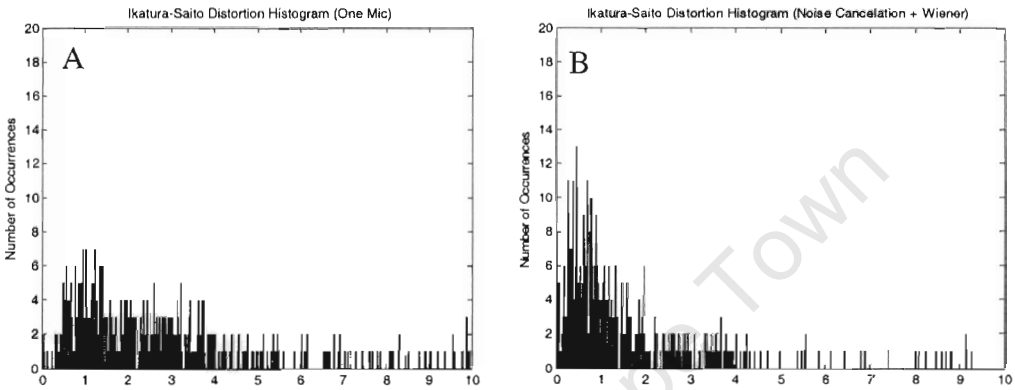

Figure 6-2: Histograms of frame-based Itakura-Saito (IS) distortion measures. (A) Baseline (single microphone) speech, and (B) Beamformed speech using NC-GSC + Wiener

Figure 6-3 clearly illustrates the difference in the levels of distortion for single microphone speech (baseline) and beamformed speech. The baseline has a higher mean and higher distortion values than the beamformed speech which has a higher concentration of values below its mean. Figure 6-3 also demonstrates that the impact of noise on degraded speech is non-uniform [65, 66]. The Itakura-Saito distortion measure shows the level of distortion for each frame across time.

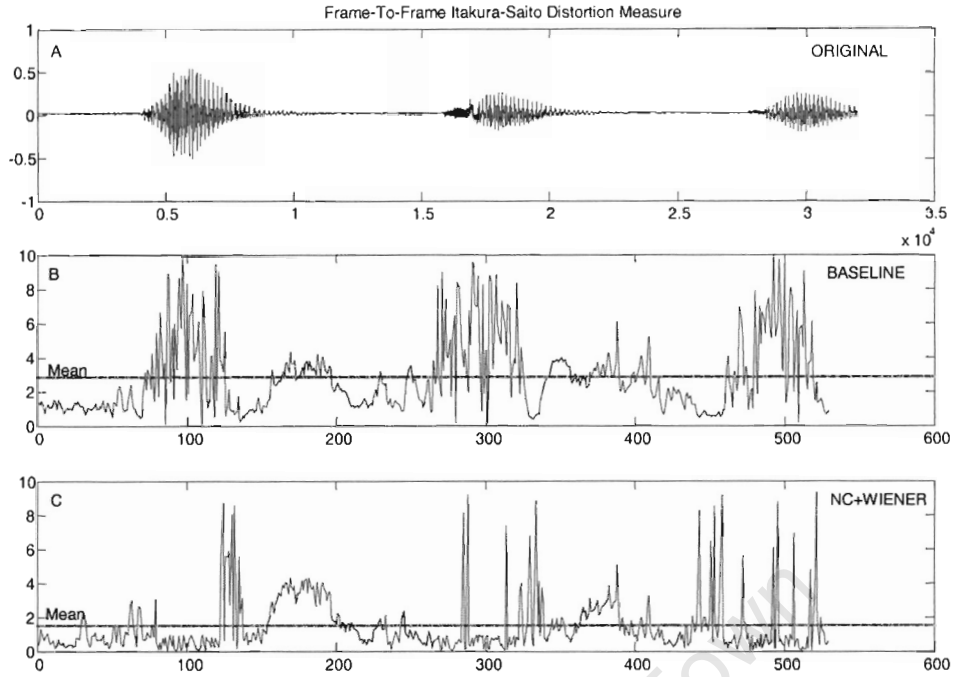

Figure 6-3: Speech waveforms of (A) original speech, (B) IS quality measure for baseline and (C) IS quality measure for NC-GSC+Wiener beamformed speech

6.1.2 Segmental signal-to-noise ratio measure

Segmental signal-to-noise ratio is a frame-based signal-to-noise ratio that is estimated by averaging frame level SNR estimates as follows,

$$d_{SegSNR} = \frac{10}{M}\sum_{m=0}^{M}\log\frac{\sum_{n=Nm}^{Nm+N-1} s_{\phi}^{2}(n)}{\sum_{n=Nm}^{Nm+N-1}\left[s_{d}(n)-s_{\phi}(n)\right]^{2}} \tag{6.2}$$

where N denotes the segment length, M denotes the number of segments, s_{ϕ} denotes the clean speech and s_{ϕ} denotes the processed speech. Frames with SNRs above 35dB do not show large perceptual differences, and generally can be replaced with 35dB in equation (6.2) [64].

Figure 6-4 is a plot of average segmental signal-to-noise ratios for 10 random speakers from the TIMIT database for four beamformers. The most notable observation is the improvement in SegSNR due to the addition of a post filter to the noise canceling beamformer.

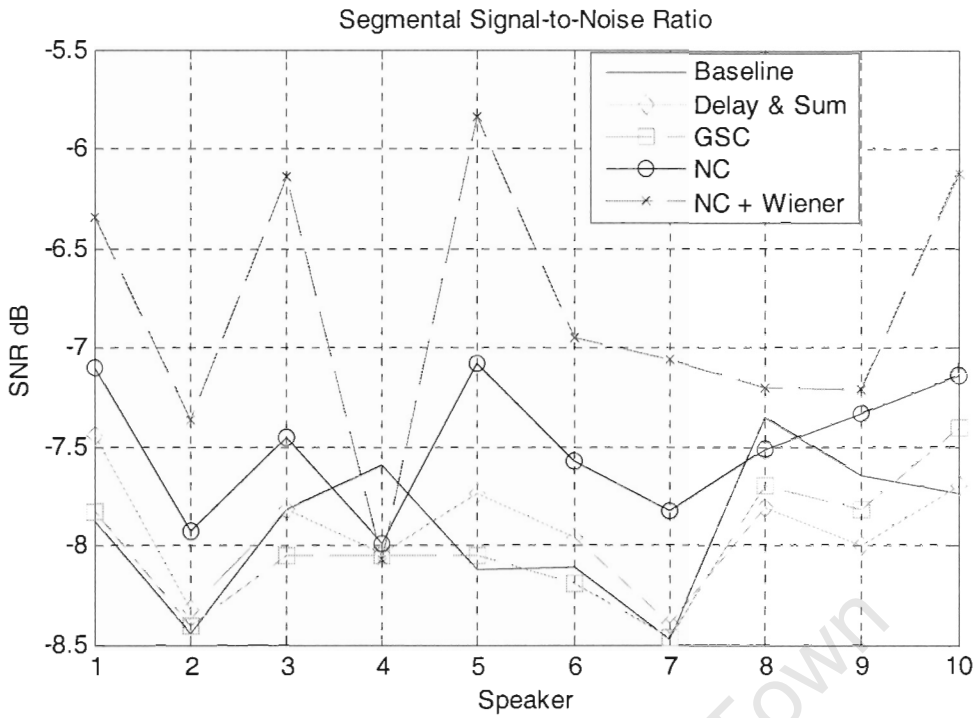

Figure 6-4: Overall mean Segmental SNR for 10 speakers

It has been shown that for clean speech recorded using a close-talking microphone, a GMM based speaker identification system similar to the one used here obtains a 100% identification rate [29]. It should be noted that the experimental setup and data used in [29] were different to that used in our evaluation. The baseline for the experiments to which all the beamforming algorithms are compared is the identification rate obtained using a single microphone under the same conditions as the microphone array. Table 6-2 below shows speaker identification and speaker verification results for 100 speakers.

6.2 Speaker Recognition Performance

An evaluation was performed on the TIMIT database to examine the effect of microphone arrays on speaker recognition systems. The TIMIT database contains a total of 6300 sentences, 10 sentences spoken by each of the 630 speakers from 8 major dialect regions of the United States. 100 speakers from the first two dialect regions in the training portion of the database were used in this evaluation. Each speaker had 10 speech segments; the first 8 segments (totaling approximately 24 seconds) were extracted and concatenated for speaker model training, and the

remaining 2 segments (totaling about 6 seconds) were concatenated and used for testing. The database of 100 speakers used consisted of 37 female speakers and 63 male speakers.

Table 6-2 summarizes the results obtained for evaluation on speaker identification and speaker verification systems. The noise canceling beamformer with a Wiener filter is successful in reducing the level of noise in the input signal, recording the highest identification rates and lowest equal error rates for distances of 50cm and 100cm from the array. As the distance of the speaker from the array increases, the performance of the microphone array system degrades.

Table 6-2: Summary of SID and SV measures for single microphone and four beamforming techniques (average ± standard error of mean)

BEAMFORMER	SID (identification rate)		SV (equal error rate)	
	50cm	100cm	50cm	100cm
Single Mic.	53%	6%	10.13%	24.26%
Delay-Sum	54%	11%	10.45%	24.56%
GSC	62%	15%	10.22%	20.55%
NC-GSC	63%	16%	9.95%	20.42%
NC-GSC+Wiener	65%	22%	8.46%	20.34%

6.2.1 Speaker identification performance

This section evaluates the beamforming algorithms previously discussed on a speaker identification task. Table 6-3 displays the performance of the beamforming algorithms on the full 100 speaker database extracted from the TIMIT database. Table 6-3 clearly illustrates that NC-GSC+Wiener outperforms the other three beamforming techniques and that the GSC beamformer performs better than the delay-sum. This result is expected, as NC-GSC is an extension of the GSC. These results emphasize the ability of the noise canceling beamformer with post-filter to compensate for noise and distortions attributed to the recording environment. The trend of the results obtained in this section corresponds to those reported in [67] in which multi-channel enhancement was carried out using adaptive noise cancellation and delay-and-sum

beamforming. However, in [67] the system was artificially implemented through simulation and the database used was degraded with synthetic noise. It should be noted that clean speech recorded using a close-talking microphone achieves a 100% identification rate when evaluated using a GMM based speaker identification system [29] similar to that used in this thesis.

Table 6-3: Effect of beamforming techniques on speaker identification performance

BEAMFORMER	SID (identification rate)			
	50cm	Relative Improvement	100cm	Relative Improvement
Single Mic.	53%	-	6%	-
Delay-Sum	54%	1.88%	11%	83.33%
GSC	62%	16.98%	15%	150.00%
NC-GSC	63%	18.87%	16%	166.67%
NC-GSC+Wiener	65%	22.64%	22%	266.67%

6.2.2 Speaker verification performance

A speaker verification system needs to either accept or reject an identity claim. As such, the system can make two types of errors, i.e., it can either falsely accept imposters or falsely reject legitimate speakers [68]. Both types of errors depend on the decision threshold used in the decision making process [69]. If the threshold is too low, the system accepts the majority of identity claims, thus making few rejections and many false acceptances. Alternatively, if the threshold is too high, the system rejects the majority of the identity, thus making few false acceptances but many false rejections. The probability of accepting a speaker given that he (or she) is an imposter is termed the *false accept rate* (FAR, or the *false alarm probability*) and is given by

$$FAR[\%] = 100 \cdot \frac{N_{FA}}{N_I} \tag{6.3}$$

where N_I is the number of imposter trials and N_{FA} is the number of times where the imposter was falsely accepted. Similarly, the probability of rejecting a speaker given that he (or she) is indeed a legitimate speaker is termed the *false reject rate* (FRR, or the *miss probability*) and is given by

$$FRR[\%] = 100 \cdot \frac{N_{FR}}{N_L} \tag{6.4}$$

where N_L is the number of legitimate speaker trials and N_{FR} is the number of times where a legitimate speaker was falsely rejected. It should be noted that the FRR can only be decreased at the expense of increasing the FAR and vice versa and depending on the application, more emphasis may be placed on one error over the other. In evaluating the performance of a speaker verification system *Detection Error Trade-off* (DET) curves [31, 69, 70] are often used. This curve plots the FAR versus the FRR using the normal deviate scale. The point on the curve where the FAR is equal to the FRR is known as the *equal error rate* (EER) and is often used as a single performance indicator for these two types of error [26]. The closer the EER value is to zero the better the performance of the system.

Table 6-4 shows the results obtained when a speaker verification system is used to evaluate the performance of the previously discussed beamforming techniques. It can be seen from the table that the NC-GSC+Wiener beamformer has the lowest EER. Figure 6-5 illustrates the plots of the DET curves for single microphone (baseline) performance and the performance of three beamforming techniques for distances of 50cm and 100cm from the microphone array. It is clearly seen that NC-GSC+Wiener performs the best with equal rates of 8.46% and 20.34% for distances of 50cm and 100cm respectively.

Table 6-4: Effect of beamforming techniques on speaker verification performance (average ± standard error of mean)

BEAMFORMER	SV (equal error rate)	
	50cm	100cm
Single Mic.	10.13 ± 0.02%	24.26 ± 0.07%
Delay-Sum	10.45 ± 0.04%	24.56 ± 0.05%
GSC	10.22 ± 0.02%	20.55 ± 0.06%
NC-GSC	9.95 ± 0.03%	20.42 ± 0.02%
NC-GSC+Wiener	8.46 ± 0.03%	20.34 ± 0.05%

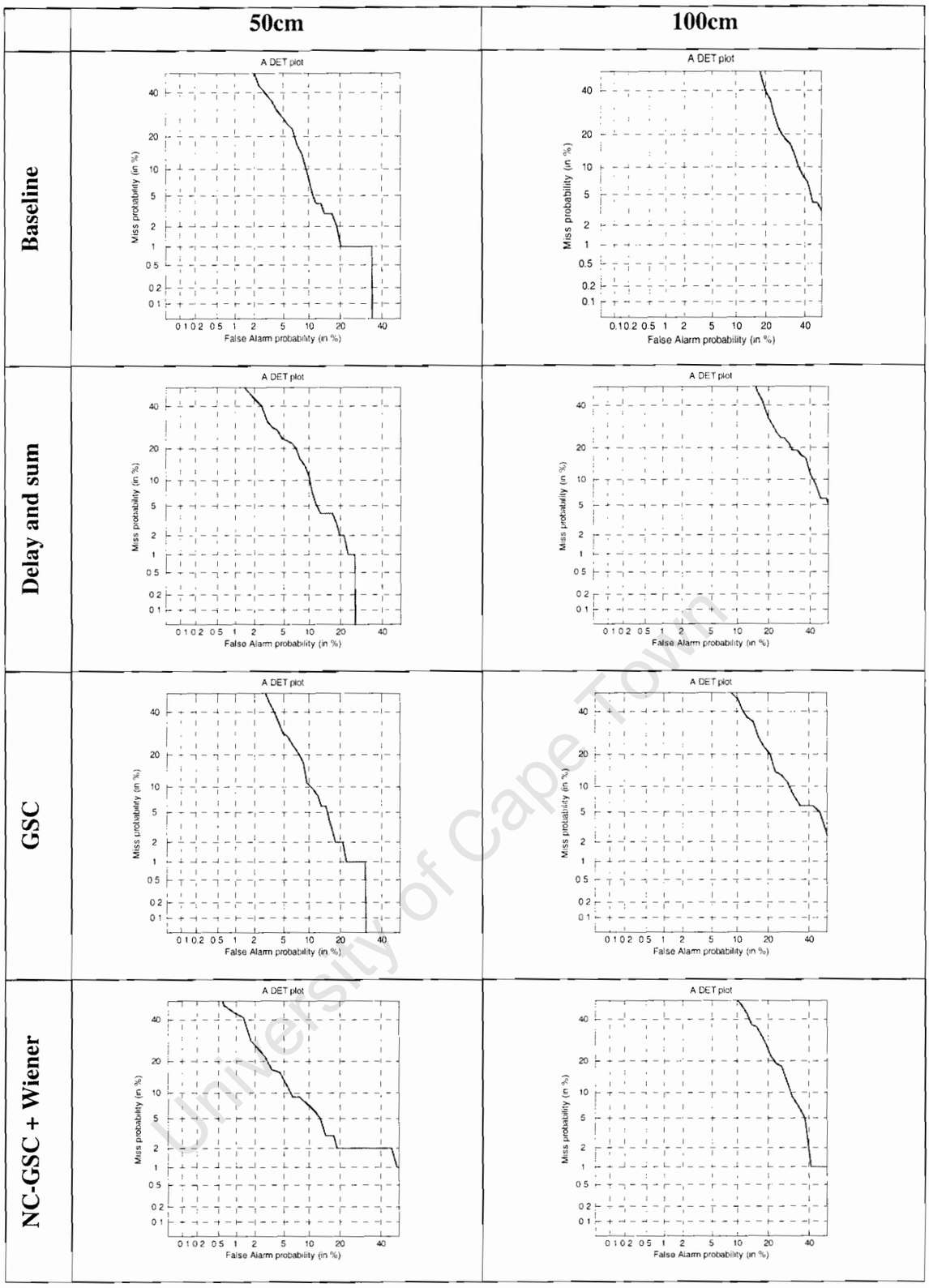

Figure 6-5: DET curves for the baseline system and three beamforming techniques

6.3 Summary

This chapter experimentally evaluates the performance of the NC-GSC+Wiener beamformer proposed in this research. The experiments conducted verify that the technique does indeed reduce noise and signal distortion in the speech input to speaker recognition systems. Furthermore, this noise canceling technique outperforms the other beamforming techniques considered in this study. NC-GSC+Wiener beamforming increases speaker identification rate by 22.64% and reduces speaker verification equal error rate by 16.49% relative to single microphone performance, for a speaker positioned 50cm from the array. For speakers positioned 100cm from the array, the NC-GSC+Wiener beamformer shows a relative increase of 266.67% (from 6% to 22%) for speaker identification and a relative reduction in equal error rate of 16.16% (from 24.26% to 20.34%). The following chapter presents a summary of the achievements of this study and provides conclusions based on the research and experimental work done.

Chapter 7

7 Conclusions

This chapter presents a brief summary and conclusions drawn from the research as described in earlier chapters. For completion, directions for future work are proposed.

7.1 Summary of work done

This thesis has investigated the use of microphone arrays for purposes of improving the robustness of hands-free speaker recognition applications in distant talking environments. In chapter 1 the objectives of this research are presented.

The first objective was to provide a comprehensive review of existing microphone array texts with emphasis on how microphone arrays work, factors affecting their performance and beamforming algorithms previously used for speaker recognition systems. This objective is attained in chapters 2 and 3. Chapter 2 specifically reviews array processing principles and discusses some key features of discrete sensor arrays while chapter 3 discussed speaker recognition and the role that microphone arrays play in speaker recognition systems.

The second objective, which was to design and implement a microphone array system using existing techniques, is discussed in chapters 4 and 5.

The final objective of the thesis was to implement the NC-GSC+Wiener beamformer and to evaluate its performance against the beamforming techniques described in

chapter 2. This objective is attained in chapters 4 and 5 where background and principles of the NC-GSC+Wiener beamformer are presented and the actual system is implemented. Chapter 6 provides an evaluation of the NC-GSC+Wiener beamformer as well as an analysis of the various results obtained. The next section highlights key conclusions based on the work done in this study.

7.2 Conclusions

The following conclusions are based on the research and experiments carried out in this thesis.

- When in a noisy, multiple source environment and without the use of a close-talking microphone, microphone arrays provide an alternative form of speech acquisition that helps improve signal distortion, signal-to-noise ratio and speaker recognition when compared with single microphone performance under similar conditions. Thus signal acquisition compensation can be regarded as an essential step in obtaining good speaker recognition performance.

- The use of a noise canceling beamformer integrated with a Wiener post-filter as presented in this research shows a substantial improvement in reducing signal distortion for an environment consistent with multiple signal sources. This is because the NC-GSC+Wiener beamformer minimizes noise levels in the signal to values close to zero while maintaining the levels of speech in the signal close to those in the original signal.

- With regard to segmental signal-to-noise ratios the NC-GSC+Wiener beamformer performs better than other beamforming algorithms discussed in this study. This is due to the ability of the NC-GSC+Wiener beamformer to minimize the noise power in each channel before summing and finally minimizing the noise power in the beamformed signal using the wiener post-filter.

- NC-GSC+Wiener beamforming can be used to improve the performance of speaker identification and speaker verification in a noisy, multiple source office environment. This is mainly due to the beamformer's ability to spatially emphasize the desired signal and reduce the levels of surrounding noise in the input signal to the speaker recognition systems. Furthermore, the NC-GSC+Wiener beamformer reduces levels of distortion in the input signal thus providing the speaker recognition front-end with normalized feature distributions.

7.3 Directions for future work

On the basis of the results and conclusions of the study performed, the following recommendations for future work are proposed:

- While some progress has been made with the NC-GSC+Wiener beamformer there is a need for further research into the development and use of more sophisticated speech enhancement techniques for speech technologies in general. Further research should be conducted into the application of the NC-GSC+Wiener beamformer on other speech technologies.

- Objective quality measures were used to evaluate the NC-GSC+Wiener beamformer in order to rate its performance against other beamforming techniques. It is important that consistent evaluations that allow proper benchmarking between different beamforming techniques are established. Standardized databases that include real world examples should be used in evaluating these techniques.

- Research into the development of better systems that incorporate different speech enhancement techniques should be explored in order to create hybrid systems that would cater for all or most of the shortfalls that current speech technology systems encounter. These systems should be developed

to counter the effects of, among other things, speaker distance, movements, different noise types, multiple speaker scenarios and reverberation effects.

In spite of all the research undertaken on speech signals and issues surrounding speech enhancement over the decades, the seemingly simple task of removing noise remains a formidable challenge to date. While it might be too early to draw conclusions, microphone arrays appear to offer an appropriate and powerful tool to advance this challenge. Although nothing can replace close-talking microphones, microphone arrays offer an alternative for speech acquisition that allows for free movement and has a great potential for minimizing the dependence of the performance of speech technologies on the recording environment. Microphone arrays are an emerging technology and as such considerable research is still being carried out in order to create systems that are more robust and that can provide more applicability and convenience in their use.

Bibliography

[1] M. L. Seltzer and B. Raj, "Calibration of Microphone Arrays for Improved Speech Recognition," *Eurospeech*, vol. 1, pp. 1005-1008, 2001.

[2] D.V. Rabinkin, R.J. Renomeron, J. C. French, and J. L. Flanagan, "Optimum microphone placement for array sound capture," presented at SPIE, 1997.

[3] M.L. Seltzer, B. Raj, and R. M. Stern, "Speech recognizer-based microphone array processing for robust hands-free speech recognition," presented at IEEE Conf. on Acoustics, Speech and Sig. Proc., Orlando, Florida, 2002.

[4] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, 1st ed: Springer, 2001.

[5] M. Omologo, M. Matassoni, P. Svaizer, and D. Giuliani, "Microphone array based speech recognition with different talker-array positions," presented at ICASSP, Seattle, Washington, 1998.

[6] I. A. McCowan, "Robust speech recognition using microphone arrays," in *Electrical Engineering*, PhD Thesis: Queensland University of Technology, Australia, 2001.

[7] Q.Lin, E. Jan, and J. Flanagan, "Microphone arrays and speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 622 - 629, 1994.

[8] J. Flanagan, A. Surendran, and E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Communication*, vol. 13, pp. 207-222, 1993.

[9] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP*, vol. 5, pp. 4-24, 1988.

[10] J. Fernandez, E. Lleida, and E. Masgrau, "Microphone array design for robust speech acquisition and recognition," presented at EuroSpeech, Budapest, Hungary, 1999.

[11] V. C. Raykar, "A study of various beamforming techniques and implementation of the constrained least mean squares (LMS) algorithm for beamforming," presented at ICASSP, Salt Lake City, 2001.

[12] L. J. Ziomek, *Fundamentals of Acoustic Field Theory and Space-Time Signal Processing*: CRC Press, 1995.

[13] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*: Prentice Hall, 1993.

[14] D. C. Moore, "Speech enhancement using microphone arrays," Masters Thesis: Queensland University of Technology, 2000.

[15] B. D. Steinberg, *Principles of Aperture and Array System Design*: John Wiley and Sons, 1976.

[16] E. Ifeachor and B. Jervis, *Digital Signal Processing: A Practical Approach*: Addison-Wesley, 1996.

[17] D. Ward, "Theory and Application of Broadband Frequency Invariant Beamforming," in *Department of Electrical Engineering*, PhD Thesis: Australian National University, 1996.

[18] I. A. McCowan, J. Pelecanos, and S. Sridharan, "Robust speaker recognition using microphone arrays," *In Proceedings of 2001: Speaker Odyssey*, 2001.

[19] H. E. Ewalt, "Speech Signal Enhancement Using A Microphone Array," in *Department of Electrical and Computer Engineering*, Masters Thesis. Milwaukee: Marquette University, 2002.

[20] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *Acoust. Soc. Am.*, vol. 78, pp. 1508-1518, 1985.

[21] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas and Propagation*, vol. 30(1), pp. 27 - 34, 1982.

[22] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 240-259, May 1998.

[23] I. A. McCowan and H. Bourlard, "Microphone array post-filter for diffuse noise field," *IEEE Int. Conf. on Acoust. Speech and Signal Processing*, vol. 1, pp. 905-908, 2002.

[24] R. L. Bouquin and G. Faucon, "Using the coherence function for noise reduction," *IEE Proceedings*, vol. 139, pp. 276-280, June 1992.

[25] J. Bitzer, K. Kammeyer, and K. U. Simmer, "An alternative implementation of the superdirective beamformer," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 991-004, October 1999.

[26] M. Skosan, "Histogram Equilisation for Robust Text-independent Speaker Verification in Telephone Environments," in *Department of Electrical Engineering*, Masters Thesis: University of Cape Town, 2005.

[27] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical Pattern Recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 4-37, 2000.

[28] H. Gish and M. Schmit, "Text-Independent Speaker Identification," *IEEE Signal Processing Magazine*, pp. 18 -32, 1994.

[29] D. A. Reynolds, "Large Population Speaker Identification Using Clean and Telephone Speech," *IEEE Signal Processing Letters*, vol. 2, pp. 46-48, 1995.

[30] M. Skosan, "Improving speaker identification performance for telephone-based applications," *SATNAC 2004*, vol. 2, pp. 135-139, 2004.

[31] G. R. Doddington, M. A. Pryzbocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, pp. 225-254, 2000.

[32] D. A. Reynolds, "An overview of automatic speaker recognition technology," *Proceedings of ICASSP 2002*, vol. 4, pp. 4072-4075, 2002.

[33] J. A. Markowitz, "Voice Biometrics," *Communications of the ACM*, vol. 43, pp. 66-73, 2000.

[34] J. M. Naik, "Speaker verification: A tutorial," *IEEE Communications Magazine*, pp. 42-48, 1990.

[35] I. A. McCowan, C. Marro, and L. Mauuary, "Robust Speech Recognition Using Near-Field Superdirective Beamforming with Post-Filtering," *Proceedings of ICASSP 2000*, pp. 1723-1726, 2000.

[36] J. Gonzalez-Rodriguez, J. Ortega-Garcia, C. Martin, and L. Hernandez, "Increasing robustness in gmm speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays," *In Proceedings of ICSLP '96*, vol. 3, pp. 1333-1336, 1996.

[37] D. A. Reynolds, "Automatic speaker recognition using Gaussian mixture speaker models," *The Lincoln Laboratory Journal*, vol. 8, pp. 173-192, 1995.

[38] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE transactions on Speech and Audio Processing*, vol. 3, pp. 72-83, 1995.

[39] I. A. McCowan, D. Moore, and S. Sridharan, "Speech enhancement using near-field superdirectivity with an adaptive sidelobe canceller and post-filter," *International Conference on Speech Science and Technology*, vol. 1, pp. 268-273, 2000.

[40] N. Wiener, *Extrapolation, interpolation and smoothing of stationary time series, with engineering applications*. New York: Wiley, 1949.

[41] H. Bode and C. Shannon, "A simplified derivation of linear least squares smoothing and prediction theory," *IRE*, vol. 38, pp. 417-425, 1950.

[42] R. Kalman, "On the general theory of control," presented at First IFAC Congress, London: Butterworth, 1960.

[43] R. Kalman and R. Bucy, "New results in linear filtering and prediction theory," *Trans. ASME, Ser. D. J. Basic Eng.*, vol. 83, pp. 95-107, 1961.

[44] T. Kailath, "A view of three decades of linear filtering theory," *IEEE Trans. Inf. Theory*, vol. IT-20, pp. 145-181, 1974.

[45] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*: Prentice-Hall, 1985.

[46] M. W. Hoffman, T. D. Trine, K. M. Buckley, and D. J. V. Tasell, "Robust adaptive microphone array processing for hearing aids: Realistic speech enhancement," *J. Acoust. Soc. Am.*, vol. 96, pp. 759-770, 1994.

[47] J. E. Greenberg and P. M. Zurek, "Preventing reverberation-induced target cancellation in adaptive-array hearing aids," *J. Acoust. Soc. Am.*, vol. 95, pp. 2990-2991, 1994.

[48] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 35, pp. 1365-1375, 1987.

[49] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, pp. 926-935, 1972.

[50] R. A. Monzingo and T. W. Miller, *Introduction to adaptive arrays*. New York: John Wiley and Sons, 1980.

[51] E. N. Gilbert and S. P. Morgan, "Optimum design of directive antenna arrays subject to random variations," *Bell Syst. Tech. Journal*, pp. 637-663, 1955.

[52] K. M. Buckley, "Broadband beamforming and the generalized sidelobe canceller," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 34, pp. 1322-1323, 1986.

[53] M. H. Er and A. Cantoni, "Transformation of linearly constrained broadband processors to unconstrained partitioned form," *IEE Proceedings Pt. H*, vol. 133, pp. 209-212, 1986.

[54] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," *ICASSP-88*, vol. 5, pp. 2578-2581, 1988.

[55] J. Bitzer, K. U. Simmer, and K. D. Kammeyer, "Multi-microphone noise reduction by post-filter and superdirective beamformer," *Proc. Int. Workshop Acoust. Echo and Noise Control*, pp. 100-103, 1999.

[56] S. A. A. Video, "A Guide to Microphone Specifications," Available at: http://www.acoustics.salford.ac.uk/acoustics_world/id/Microphones/Microphones.htm [Last Accessed: 31/08/2005], 2005.

[57] M. Pollick, "What are Omnidirectional Microphones?" Available at: http://www.wisegeek.com/what-are-omnidirectional-microphones.htm [Last accessed: 31/08/2005], 2005.
[58] S. Devices, "Microphone Python Powering," Available at: http://www.sounddevices.com/tech/phantom.htm [Last accessed: 21/09/2005], 2005.
[59] N. S. Corporation, "LM386 Low Voltage Audio Power Amplifier," 2000.
[60] E. Technology, "PCI PnP Analog Board User's Manual," 2001-2002.
[61] W. f. Windows, "WaveView for Windows User's Manual," 2003.
[62] iotech, "WaveView: Out-of-the-Box Software," Available at: http://www.iotech.com/catalog/software/waveview.html [Last accessed: 10/10/2005], 2005.
[63] D. Schreck and S. Nelson, "Directional Microphone Array Processing Unit," Stevens Institute of Technology, Hoboken April 24 1998.
[64] J. H. L. Hansen and B. L. Pellom, "An Effective Quality Evaluation Protocol for Speech Processing Algorithms," *ICSLP*, vol. 7, pp. 2819-2822, 1998.
[65] J. H. L. Hansen and M. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 4, pp. 795-805, 1991.
[66] S. Nandkumar and J. H. L. Hansen, "Dual-channel iterative speech enhancement with constraints based on an auditory spectrum," *IEEE Trans. SAP*, vol. 1, pp. 22-34, 1995.
[67] J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Overview of speech enhancement techniques for automatic speaker recognition," *ICSLP 1996*, vol. 2, pp. 929-932, 1996.
[68] J. Koolwaaij, *Automatic speaker verification in telephony: A probabilistic approach*: PrintPartners Iskamp B. V., Enschede, 2000.
[69] F. Bimbot, J. Bonastre, C. Fredouille, G. Gravier, I. Margin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430-451, 2004.
[70] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," *EuroSpeech 1997*, pp. 1895-1898, 1997.