2011

In air acoustic vector sensors for capturing and processing of speech signals

Muawiyath Shujau
University of Wollongong

Recommended Citation

Shujau, Muawiyath, In air acoustic vector sensors for capturing and processing of speech signals, Doctor of Philosophy thesis, School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, 2011. http://ro.uow.edu.au/theses/3454

In Air Acoustic Vector Sensors for Capturing and Processing of Speech Signals

A Thesis submitted in (partial) fulfilment of the requirement for the award of the degree

Doctor of Philosophy

from

UNIVERSITY OF WOLLONGONG

By

Muawiyath Shujau
Bachelor of Engineering (Honours I)

School of Electrical, Computer and Telecommunications Engineering

August 2011

Abstract

Capturing speech signals for enhancement is an important stage in all modern communication systems. Traditionally, speech enhancement is performed on a single channel recording, but recently the advantages of multichannel speech processing have been indentified. The multichannel speech signals are captured using a microphone array, and by using the spatio-temporal information at the output of the microphone array the directional information of the source can be derived and spatial filtering of the captured signal can be performed, which show superior performance over single channel approaches. Generally, spatially distributed microphone arrays as used in speech signal processing, only capture the acoustic pressure. In this thesis, however, a co-located microphone array which captures both acoustic pressure and particle velocity, known as an Acoustic Vector Sensor (AVS), will be used for capturing speech signals for enhancement.

The AVS used in this work consists of two pressure gradient sensors and an omni-directional microphone which enables the capturing of speech of signals in 2D. Compared with other microphone arrays, the size of the AVS array is small, occupying a volume of approximately 1cm^3. The small size of the AVS array enables it be used in mobile electronic devices such as mobile phones and mobile personal computers which traditionally have a single microphone capsule.

In this thesis, a design change for the AVS is presented, which, improves the accuracy of Direction of Arrival (DOA) estimates from the AVS. It is shown that by offsetting the directional sensors on the AVS array, a source direction can be identified with an accuracy of two degrees for a stationery speech source and five degrees for both moving and multiple speech sources. Here, DOA estimates are found using the MUltiple SIgnal Classification (MUSIC) Algorithm in the time domain and an intensity based algorithm in the frequency domain. For multiple sources, a new data clustering technique is introduced with the existing frequency domain intensity based algorithm.

Speech enhancement methods, which take advantage of the directional characteristics of the AVS array are presented. It is shown that by taking advantage of the directional characteristics of the AVS to obtain noise estimates used in the Minimum Variance Distortionless Response (MVDR) beamformer, an improvement of

1.34 Mean Opinion Score (MOS) was achieved over the conventional MVDR beamformer. Here, the noise covariance matrix is obtained by a new technique which uses Singular Value Decomposition (SVD) of the AVS array outputs. Furthermore, it is shown that by applying the Griffiths and Jim (GJ) beamformer to the AVS output channels, a MOS of 1.74 over unprocessed noise corrupted speech signals was achieved in listening tests.

A new technique for speech enhancement which combines Linear Predictive (LP) spectrum-based perceptual filtering to the recordings obtained from an AVS is presented. The technique takes advantage of the directional polar responses of the AVS to obtain a significantly more accurate representation of the LP spectrum of a target speech signal in the presence of noise when compared to single channel, omni-directional recordings. Listening tests results show significant improvements in MOS scores of 1.6 over unprocessed noise corrupted speech. Further improvements to the proposed LP spectrum based perceptual filtering are achieved by introducing the averaged autocorrelation function to obtain a multichannel LP spectrum from the directional components of the AVS array. By introducing the average autocorrelation function a MOS of 1.98 over unprocessed noise corrupted speech signals is achieved.

In addition to the perceptual filter, two Blind Source Separation (BSS) algorithms are presented. The well known Independent Component Analysis (ICA) and a new method based on the clustering of DOA estimates performed on a time frequency basis are presented. Comparisons are made between co-located microphone arrays that contain microphones with mixed polar responses and traditional Uniform Linear Arrays (ULA) formed from omni-directional microphones and Soundfield microphones. It is shown that polar responses of the microphones are a key factor in the performance of ICA applied to co-located microphones. It is shown by applying the two BSS algorithms, improvements of 1.75 and 2.09 MOS over unprocessed noise corrupted speech signals are achieved for ICA and DOA based methods respectively, during listening tests.

Finally, the DOA estimation and clustering method for BSS is used for dereverberation of speech signals. It is shown that by using the directional characteristics of the AVS array, reflections from different directions can be minimized. The results show that an improvement in terms of Signal to Reverberant Ratio (SRR) of

1.5 dB and 2.5 dB for a source at 1m and 5m from the AVS array respectively is achieved.

Thesis Certification

I Muawiyath Shujau, declare that this thesis, submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Electrical, computer and Telecommunications Engineering, University of Wollongong, is wholly my work unless otherwise referenced or acknowledged. The document has not been submitted for qualifications at any other academic institution.

Muawiyath Shujau

25 August 2011

Acknowledgements

I would like to thank my supervisors Dr. Christian Ritz and Prof. Ian Burnett for all the help, support and guidance they have given me throughout my research, without which I would have not been able to complete my work.

To my mother and father who endured unimaginable hardships to get me to this point in my life, I thank them whole heartedly for all the love, support and prayers.

To my wife Shaira, for the love, support, patience and sacrifice during this three and half years, without which I would have not been able to complete this work.

To my beautiful daughter Mazaya, my inspiration for pursuing this degree, I hope I make you proud.

And finally to all my family and friends for all the help and support, I dedicate this work to you all.

Contents

List of Figures

List of Tables

List of Abbreviations

AR	Auto Regression
ATF	Acoustic Transfer Function
AVS	Acoustic Vector Sensor
BL	Back Left
BR	Back Right
BSS	Blind Source Separation
DF	Distance Factor
DMOS	Degradation Mean Opinion Score
DOA	Direction of Arrival
DS	Delay and Sum
DUET	Degenerate Un-mixing Estimation Technique
DYPSA	Dynamic Programming Phase Slope Algorithm
ESPRIT	Estimation of Signal Parameters via Rotational Invariance Technique
ESS	Exponential Sine Sweep
FET	Field Effect Transistor
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FL	Front Left
FR	Front Right
GCC	Generalized Cross Correlation
GCC PHAT	Generalized Cross Correlation with Phase Transform
GCI	Glottal Closure Instances
GJ	Griffiths and Jim
GSC	Generalized Sidelobe Canceller
ICA	Independent Component Analysis
ISD	Itakura Saito Distance
ITU	International Telecommunications Union
JADE	J. F. Cardoso’s ICA algorithm
LCMV	Linearly Constrained Minimum Variance
LPC	Linear Predictive Coding

LP	Linear Prediction
LQO	Listening Quality
LSD	Log Spectral Distortion
ML	Maximum Likelihood
MMSE	Minimum Mean Square Error
MOS	Mean Opinion Score
MOS LQO	Mean Opinion Score Listening Quality
MSC	Multiple Sidelobe Canceller
MSE	Mean Square Error
MSNR	Maximization of the Signal to Noise Ratio
MUSHRA	Multi-Stimulus test with Hidden Reference and Anchor
MUSIC	MUltiple SIgnal Classification
MVDR	Minimum Variance Distortionless Response
MV	Minimum Variance
PDF	Probability Density Function
PESQ	Perceptual Evaluation of Speech Quality
PESQ MOS	Perceptual Evaluation of Speech Quality Mean Opinion Score
PHAT	Phase Transform
PSD	Power Spectral Density
RE	Random Efficiency
RT_{60}	Reverberation Time
SDR	Signal to Distortion Ratio
SIR	Signal to Interference Ratio
SNR	Signal to Noise Ratio
SMERSH	Spatiotemporal Averaging Method for Enhancement of Reverberant Speech
SRP	Steered Power Response
SRP PHAT	Steered Power Response with Phase Transform
SRR	Signal to Reverberation Ratio
SVD	Singular Value Decomposition
TDE	Time Delay Estimate
TDOA	Time Difference of Arrival
TIMIT	Texas Instruments/Massachusetts Institute of Technology
ULA	Uniform Linear Array

VAD	Voice Activity Detection

Chapter 1 Introduction

1.1 Overview

In the past two decades, demand for efficient and high quality speech signal processing tools and algorithms have been increasing. The increase in demand is due to the increase in popularity of mobile devices such as mobile phones, wireless mobile computers and availability of wireless broadband access from almost any location. Applications such as teleconferencing, hands free mobile telephony, remote class rooms and remote telemedicine are some applications that require high quality speech signal processing.

Speech signals are traditionally captured using a single microphone and all processing is based on a single channel. The single channel signals lack the ability to provide a detailed description of the recording environment and it limits the ability for applications such as video teleconferencing when there is more than one user in the room. The current trend in capturing speech signals is based on using multiple microphones arranged in different orientations known as a microphone array. The multi microphone scenario facilitates the application of signal processing approaches that allows the ability to locate sources, separate individual sources when there are more than one source, enhance noise corrupted speech and it allows the capture of 3D soundfields.

The applications described above require design of high quality, compact and low cost microphone arrays. In the past, most microphone arrays were designed to take advantage of the spatial distribution of capsules such that statistically independent and time delayed recording of sources can be made. These two features of the captured signals were used in processing the signals for beamforming and speech enhancement. In this thesis, a microphone array known as an Acoustic Vector Sensor (AVS) that has all its capsules co-located is proposed. This is a unique microphone array that contains one scalar pressure sensor (omni-directional sensor) and three pressure gradient sensors (pressure gradient microphones) arranged orthogonally such that the sensors point in the x, y and z directions in 3D space. The total volume occupied by the capsules in the AVS array of this thesis is approximately 1cm^3. The pressure gradient sensors capture both the sound intensity and the particle velocity of the soundwave. The size of the proposed array compared to traditional microphone arrays is extremely small. Hence, these sensors can be used in mobile devices such as mobile phones, tablets and other small mobile computing devices. While the original application of the AVS was for sonar in water, the work presented here is targeted for in-air speech recordings.

In particular this thesis will consider signal processing of AVS speech recordings for four major application areas; speech source direction of arrival estimation, beamforming, speech enhancement and source separation. Although these four areas are treated differently, in almost all the literature there is a close relationship between them. Of the methods listed above, direction of arrival estimation and beamforming methods have been used for over fifty years and most of these algorithms have their roots in narrowband radar and sonar applications. The arrays that were used to capture signals for processing with these algorithms consisted of spatially distributed microphone capsules. Here, these algorithms will be used for signals captured from an array formed from co-located microphone capsules.

The work presented here will show the advantages of using a co-located array such as the AVS for capturing and processing speech signals. It will be shown that there are hardware features of the array that enable better performance in terms of accurate DOA estimation, beamforming and speech enhancement compared to other microphone arrays even without any processing of the signals. One of the key advantages of using an AVS is its small size, when compared to other arrays designed for 3D soundfields. The size of an AVS is not only small in terms of physical size of the array but the number of capsules used in the construction. Comparisons of the performance of the array will be made with other microphone arrays that are comparable in size and number of capsules used.

1.2 Thesis Outline

The work presented in this thesis is organised as follows: Chapter 2 presents background knowledge needed to understand the content of this thesis and a critical review of microphones and microphone array signal processing, especially for a co-located microphone array. The first part of this chapter is dedicated to the fundamentals of soundwaves which are essential for understanding the concepts that will be presented later. Microphone theory is covered in detail, with emphasis on the derivation of mathematical theory of directional microphones. This derivation of the directional characteristics is essential in the design of the AVS. A review of techniques used in Direction of Arrival (DOA) estimation is given for a general microphone array and techniques which are applicable to a co-located microphone array are highlighted. The review shows that any DOA estimation algorithm that does not rely on (Time Difference of Arrival) TDOA can be used for DOA estimation using an AVS. A detailed examination of beamforming algorithms are presented next. Here, emphasis is on the application of beamformers to a co-located microphone array. Finally, speech enhancement algorithms and performance evaluation tools for speech enhancement are presented. This review highlights the close relationship between speech enhancement, source separation and beamforming.

In Chapter 3, the design of the AVS is investigated with emphasis on improving the performance of the AVS in terms of DOA estimation. The proposed design changes to existing AVS arrays proposed in the literature will be justified by means of the measured accuracy of the DOA estimation, and mathematical reasoning will be provided to justify the changes that are made to the AVS design. Polar plots for monotone frequencies covering 1 to 10 kHz will be shown for existing AVS arrays and for the improved AVS arrays. Further improvements to the performance to account for manufacturing defects of the array will also be investigated and a solution to correct these errors in software will be presented.

Chapter 4 looks at DOA estimation for stationary and moving speech sources. A comparison between the performance of an AVS and a Soundfield microphone will be presented. Two different techniques used for DOA estimation will be presented: the well known MUltiple SIgnal Classification (MUSIC) algorithm and an intensity based algorithm that is unique to the directional co-located microphone arrays. An investigation into the size of a speech frames that can give accurate DOA estimation and the importance of Voice Activity Detector (VAD) in the DOA estimation of the speech signals are presented here. Finally, DOA estimation of multiple speech sources will be presented for both MUSIC and an intensity based algorithm.

Beamforming, speech enhancement and source separation algorithms for the AVS will be presented in Chapter 5. A database of recordings from the AVS in anechoic and reverberant conditions containing speech corrupted by different noise sources and other speech sources is presented. This database contains over 300 recordings from the AVS and is used in the evaluation of the performance of the different algorithms.

Here a perceptual based Wiener filtering approach is applied to the AVS signals, which results in high quality enhancement as judged by subjective and perceptual based objective tests. The proposed approach makes use of multichannel Linear Prediction (LP) coefficients and beamforming. The performance of the perceptual filter is compared against the Minimum Variance Distortionless Response (MVDR) Beamformer.

Beamforming algorithms, which have been used in microphone array signal processing will be applied to an AVS. The two most well known beamformers, the MVDR Beamformer and the Griffiths and Jim (GJ) beamformer, will be applied to an AVS. An extension to the MVDR beamformer based on an AVS array will be presented, which improves the speech quality of the beamformer output in noise corrupted speech.

A new source separation algorithm using intensity based DOA estimation will be presented. The algorithm presented here uses clustering techniques and binary masking based on DOA estimates applied on a time frequency basis to separate sources, in multisource scenarios. A comparison is made between the proposed algorithms and the well known ICA algorithm.

Dereverberation based on the proposed source separation techniques is presented; recordings with high reverberation times are de-reverberated using the proposed technique and the well known Spatiotemporal Averaging Method for Enhancement of Reverberant Speech (SMERSH) algorithm. Generally, most dereverberation algorithms are based on the impulse response for dereverberation, which requires estimation or prior knowledge; the algorithms presented do not used the room impulse response for dereverberation.

Finally, source separation algorithms are used in the enhancement of the noise corrupted speech signals and comparisons are made between the performance of beamformers, speech enhancement techniques and source separation techniques. Chapter 6 presents the conclusion of the thesis and summarises the major findings and identifies potential areas where this research can be expanded in the future.

1.3 Contributions

The contributions made in this work are presented below. The contributions are arranged according to the order they appear in the thesis. The contributions and the associated publication are listed.

- Improvement of the design of an AVS array to improve the accuracy of DOA estimation is presented. These design improvements enable DOA estimates of monotone and speech signals with average accuracy of approximately 4 degrees for both anechoic and reverberant recordings. (Chapter 3) [1]
- DOA estimation of speech signals from stationary and moving sources is presented, with emphasis on the relation between frame size, voiced and unvoiced regions of speech and speed of a moving source. It is shown that a frame size of 20ms is sufficient to get an accurate DOA estimate from an AVS array. A method for determining DOAs for multiple consecutive speakers is presented for two and three speakers. (Chapter 4) [2, 3]
- Different methods of beamforming for an AVS array are shown. The MVDR Beamformer and the GJ beamformer are applied to the output of an AVS. It is shown that basic assumptions made in the derivation of the MVDR Beamformer can be achieved by incorporating a stage where accurate estimates of noise and the interfering signals are obtained by using Singular Value Decomposition (SVD) decomposition on paired channels of the AVS. The noise and interference signal estimates from the SVD is then used in formation of a more accurate covariance matrix, which in turn is used in the MVDR Beamformer. The results show that there is an improvement in terms of Perceptual Evaluation of Speech Quality Mean Opinion Score (PESQ MOS) score when the modification is made to the MVDR Beamformer compared to the traditional approach. (Chapter 5)
- Speech enhancement based on a modified perceptual wiener filter is presented. Here a single channel algorithm is modified for the channels of an AVS. The key contribution here is the use of the directional features of the AVS channels to get an accurate representation of the LP spectra of the

speech signal which is used in the formation of the perceptual filter. (Chapter 5) [4]

- Speech enhancement for noise corrupted speech signals based on the Independent Component Analysis (ICA) algorithm applied to an AVS is presented. The ICA algorithm normally works on spatially distributed microphone channels. Here it is shown that due to the directionality of AVS channels, the statistics of channels are independent enough for the basic assumptions made in ICA to be fulfilled. Hence, ICA can be applied to the AVS channels directly. The results show improvements in speech quality in terms of PESQ scores. (Chapter 5) [5]
- A source separation algorithm using intensity based DOA estimation approach is presented. The key features of this algorithm is its use in the sorting of DOA estimations to form individual sources and the use of binary masking to separate the frequency components of individual sources based on the sorted DOA estimations. The results from listening tests and Signal to Interference Ratio (SIR) and Signal to Distortion Ratio (SDR) show good performance of the proposed algorithm compared to the well known ICA algorithm. (Chapter 5) [3]
- The different method for obtaining the accurate estimate of Linear Prediction (LP) spectra is shown. The enhancement techniques for AVS that was discussed before are used, in addition to these algorithms multichannel LP spectra are incorporated into the algorithm and different methods for obtaining multichannel LP spectra are investigated. (Chapter 5)
- The directional characteristics of the AVS channels are tested for their use in dereverberation. It is found that compared to omni-directional sensors the directional sensors produce less reverberant recordings. A dereverberation algorithm based on DOA estimates is presented. The algorithm is similar to the source separation algorithm presented before and comparisons are made against the well known SMERSH algorithm. Results presented show, in highly reverberant conditions the proposed method outperforms the SMERSH algorithm. (Chapter 5) [3]

1.4 Publications

1.4.1 Conference Publications

1. M. Shujau, C. H. Ritz, and I. S. Burnett, "Designing Acoustic Vector Sensors for localisation of sound sources in air," presented at the *17th European Signal Processing Conference (EUSIPCO 2009)*, Glasgow, Scotland., 2009.

2. M. Shujau, C. H. Ritz, and I. S. Burnett, "Speech enhancement via separation of sources from co-located microphone recordings," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 137-140.

3. M. Shujau, C. H. Ritz, and I. S. Burnett, "Using in-air Acoustic Vector Sensors for tracking moving speakers," in *Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference on*, 2010, pp. 1-5.

4. M. Shujau, C. H. Ritz, and I. S. Burnett, "Linear Predictive Perceptual Filtering For Acoustic Vector Sensors: Exploiting Directional Recordings For High Quality Speech Enhancement," *presented at the Acoustic Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, Praque, 2011*.

5. M. Shujau, C. H. Ritz, and I. S. Burnett, “Separation of Speech Sources Using An Acoustic Vector Sensor,” *accepted for presentation at the Multimedia Signal Processing (MMSP), 2011 IEEE internatinal Workshop on, Hanzhou, 2011*

1.4.2 Book Chapters

6. Christian H. Ritz, Muawiyath Shujau, Xiguang Zheng, Bin Cheng, Eva Cheng and Ian S Burnett (2011). Backward Compatible Spatialized Teleconferencing based on Squeezed Recordings, *Advances in Sound Localization*, Pawel Strumillo (Ed.), ISBN: 978-953-307-224-1, InTech, Available from: http://www.intechopen.com/articles/show/title/backward-compatible-spatialized-teleconferencing-based-on-squeezed-recordings

1.4.3 Papers to be Submitted

7. M. Shujau, C. H. Ritz, and I. S. Burnett, “Dereverberation using speech source separation based on an Acoustic vector sensor,” *for submission to Acoustic Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, Kyoto, 2012*
8. M. Shujau, C. H. Ritz, and I. S. Burnett, “DOA estimation of multiple speech source based on an In-air Acoustic vector sensor,” *for submission to Acoustic Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, Kyoto, 2012*
9. M. Shujau, C. H. Ritz, and I. S. Burnett, “Methods for obtaining an accurate LP spectra for perceptual filtering using the output channels of an AVS,” *for submission to IEEE Transactions on Signal Processing.*

Chapter 2 Literature Review

2.1 Introduction

The increased demand for capturing high quality speech signals in communication system has seen a shift from single channel recordings to multichannel recordings with microphone arrays. The multichannel recording provides much more information about the speech signals hence enabling better processing especially, in noisy and reverberant environments. There are several types of microphone arrays; the restriction in using these microphone arrays in mobile devices is the size. The physical size and the number of microphones in a standard array formed from spatially distributed microphone have to be large to take full advantage of the array, which limits their use in mobile devices. Hence, there is a high demand for a microphone array that is small in size and capable of delivering high quality recordings of speech with directional and spatial information. The work presented in this thesis will be based on such a microphone array, which is a compact and co-located known as an Acoustic Vector Sensor (AVS). The AVS is capable of measuring three orthogonal components of the particle velocity of the soundwave and the pressure signal simultaneously in the same location using three velocity gradient sensors placed orthogonally to each other and pointing in the x, y and z directional and an omni-directional sensor.

2.2 Definition of an AVS

An array capable of measuring both particle velocity and pressure of a soundwave in three dimensions at a given point in space can be described as an AVS. The size of the structure, the capsules and the arrangement in which the capsules are attached to the array, all contribute to the accuracy with which the directional information is captured by an AVS [6]. The theoretical derivation of the performance of the AVS has been shown through Cramer-Rao bounds for localizing sound sources in [7] and it has been shown that the accuracy of DOA estimation and beamforming of the AVS is better compared to other microphone arrays of comparable capsule number and size.

2.3 Origin of AVS

The first AVS was used for DOA estimation of electromagnetic waves [8]. The early version of the AVS in [8] uses two orthogonal triads of scalar sensors that measure the complete electric and magnetic field of the source at the sensors. The advantages offered by the array include capturing of all available information about the electromagnetic waves at the sensor and smaller array apertures, which increases the accuracy of the DOA estimates over conventional scalar sensors. The idea presented in [8] for electromagnetic waves is extended to the acoustic case in [7] for DOA estimates of acoustic sources in underwater applications.

The AVS array for acoustic signals in [8] is constructed using four sensors of which three are acoustic particle velocity or gradient sensors and one is an acoustic pressure sensor. The sensors are arranged such that, three gradient sensors are mounted orthogonally to each other facing the x, y and z directions in three dimensional spaces. The sensors are mounted such that the volume occupied by the sensors is minimised. The minimization of the volume is important for the assumption of co-located sensors to be valid. The vector sensors used in [7] are true acoustic particle velocity sensors, known as a hot wire anemometer which will be discussed in detail later in this chapter. An alternative method for the construction of the AVS is presented in [9], where pressure gradient sensors are used to replace the anemometers. This design is more suitable for in-air applications such as speech audio signal processing.

2.4 Applications of AVS

The early applications of the AVS array were generally in the estimation of the detection of electromagnetic waves and in underwater acoustic applications such as seismic activity detection [10] [11]. Similarly an AVS has been used in sonar applications and this is one of the major areas where AVS's are used [12-14]. The majority of research into AVS arrays is based around these major applications in underwater scenarios. The AVS has also been used in-air for detecting the movement of battle field vehicles [15]. Furthermore, in [9, 16, 17], AVS's are used for source localization of wideband sources and in noise reduction. The bulk of the applications for the AVS are not for speech; in fact little literature exists of an AVS applied to speech processing in terms of speech source localization, speech enhancement and source separation. The only available literature on the AVS for speech signals is present in [9] and [18], where AVS signals were used for source localization and two AVS arrays were used for binaural multichannel beamforming.

The rest of this chapter is organised as follows: An introduction of the properties of soundwaves will be followed by devices used in the capture of soundwaves. Then, microphone arrays and signal processing for microphone arrays in general will be presented, where DOA estimation, beamforming, speech enhancement and dereverberation will be discussed for multichannel recordings. The remainder of this chapter will describe the foundations that are needed to understand different concepts and background theoretical knowledge that is required for the work presented in this thesis.

2.5 Soundwaves

Soundwaves are waves that move due to the molecules of a fluid vibrating horizontally to the direction of propagation. These vibrations cause changes in pressure, density and temperature of the molecules in the medium. For a soundwave there are several important relationships that govern the characteristics of the wave. The most important of these relationships are those between the particle velocity, temperature, density and pressure [19, 20], which are critical for capturing the sound accurately, especially when DOA estimates of the soundwaves are needed.

2.5.1 Velocity of Sound in Air

The velocity of a soundwave is described by the relationship between the density of the material and Young's modulus as:

$$v = \sqrt{\frac{E}{\rho}} \tag{1}$$

where E is Young's modulus, ρ is the density of the material and v is the velocity of the soundwave. For air, which does not have a Young's modulus, the soundwave propagation is considered as an adiabatic process, where there is no heat transfer [20]. By using the gas laws it can be shown that an equivalent to the Young's Modulus for air can be expressed as $E_{gas} = \gamma P$ where γ is constant which depends on the gas (for air $\gamma = 1.4$). The velocity of sound in air is given as[20]:

$$v_{gas} = \sqrt{\frac{E_{gas}}{\rho_{gas}}} = \sqrt{\frac{\gamma R_{mol} T_{abs}}{M_{mass}}} \tag{2}$$

where R_{mol} is the gas constant (8.31 JK^{-1} mole^{-1}), T_{abs} is the absolute temperature (K) and M_{mass} is molecular mass of the gas (kg mole^{-1}). This relationship shows that the velocity of the soundwave is not affected by the pressure, but by the temperature and the molecular mass of the gas.

Soundwaves that carry information such as speech, behave differently from monotone signals where the frequency and amplitude of the soundwave remains constant. The speech information is contained in changes in frequency and pressure level (the amplitude of the soundwave). The human ear or microphones must be capable of detecting the changes in the pressure as well as the frequency. The relationship between the frequency and the wavelength of a soundwave in air is described as:

$$v = f\lambda \tag{3}$$

where v is velocity of sound in air, f is the frequency of the soundwave and λ is the wavelength. In this thesis, the velocity of sound is assumed to be 344 ms^{-1}. Here, the velocity of air describes the wave as a whole; another quantity that describes the velocity of the particles in the wave is the particle velocity of the soundwave.

2.5.2 Particle Velocity of Soundwave

Particle velocity of the soundwave is the velocity of the molecules as they oscillate around the origin, which is not equal to the velocity of sound. Normally the velocity of sound is much higher than that of the particle velocity of the molecules. The relation between the particle velocity and the pressure is [20]:

$$v_{par} = \frac{j}{k_0 z_0} \nabla p \tag{4}$$

where j is $\sqrt{-1}$ and signifies that the driving force leads particle velocity by $\frac{\pi}{2}$ radians. $k_0 = \frac{\omega}{c_0}$ is the free field wave number of a plane wave, $z_0 = \rho_0 c_0$ is the wave impedance of free plane wave and ∇p is the gradient of the pressure of the wave [19-21]. From this relationship it can be seen that particle velocity of the soundwave is proportional to the gradient of the sound pressure. Hence, any electro mechanical device capable of capturing the change or the gradient of the pressure between two points can capture the particle velocity component of the soundwave.

2.5.3 Intensity of the Soundwave (Energy of Soundwave)

The propagation of the soundwave thus far has been considered in only one direction. But in reality, the soundwave moves outwards from the source in all directions and spreads out as it travels further away from the source. The intensity or the energy of the soundwave at a point in space from the source is given as:

$$I = Pv_{par} = \frac{W_{source}}{A_{surface}} \tag{5}$$

where P is the sound pressure at the source, W_{source} is power of the source in Watts and $A_{surface}$ is the surface area (the surface area with which the soundwave comes in contact with) . If it is assumed that the soundwave expands out as a sphere, then the intensity of the soundwave is expressed as:

$$I = \frac{W_{source}}{4\pi r_{radius}^2} \tag{6}$$

where r_{radius} is the distance from the source. Here, the intensity of the soundwave weakens as it moves away from the source according to the inverse square law of (6). The sound intensity can vary over a large range (greater than 10^{12}) and since human beings perceive loudness on a logarithmic scale the sound intensity level is usually expressed on a logarithmic scale. The sound intensity level can be expressed as:

$$SIL = 10\log_{10}\left(\frac{I_{actual}}{I_{ref}}\right) \tag{7}$$

where I_{actual} is the actual sound power flux level and I_{ref} is the reference sound power flux. Since it is hard to measure sound intensity and human ears detect sound pressure rather than sound intensity level, a more practical measure for describing the amplitude of a soundwave is the sound pressure level.

The Sound Pressure Level (SPL) is defined as:

$$SPL = 20\log_{10}\left(\frac{P_{actual}}{P_{ref}}\right) \tag{8}$$

where P_{actual} is the actual pressure level (in Pa) and P_{ref} is the reference pressure level $(20\mu Pa)$. The reference pressure level is known as the threshold of human hearing at 1 kHz. Here, the factors 20 and 10 are the integer change that is approximately equal to the smallest change that can be perceived by the human ear.

2.5.4 Multiple Sound Sources

In most real situations there is more than one sound source present and this may be due to two individual sources or it may be due to delayed reflections of same source. Hence, there are two scenarios that have to be considered when the sound levels from different sources are combined. That is, the sources may be:

- Correlated Sources
- Uncorrelated Sources

2.5.5 Correlated Sources

Correlation means that two statistical processes are related. In terms of sound sources, this means that the sources are related to each other. This may occur if the signals from two or more loud speakers separated in space are playing the same recording or if the signal is reflected from the walls of a room with small delays. For correlated sources, the waves from different sources have the exact same frequencies and if the soundwaves are in phase they simply add producing a signal that increases in magnitude. If the signals are not in phase then the waves are added depending on the phase of the individual components, in this case the magnitude of the combined wave is less than that of the original signal.

2.5.6 Uncorrelated Sources

Uncorrelated sound sources are those that have no statistical relation between the two sources. This is the case when there is more than one person speaking at a time, or when there are different instruments been played in an orchestra. The key difference here is the frequency components of the different sources are not the same.

A signal reflected from the walls of a highly reverberant room, where the delay between the original signal and the reflected signal is high is also considered to be uncorrelated. When the soundwaves are uncorrelated they combine differently to that of the correlated sources. The power of the different waves is added together. The power of the soundwave is given by the square of the sound pressures. The combined sound pressure of the uncorrelated sources is give as [20]:

$$P_{total} = \sqrt{P_1^2 + P_2^2 + \cdots + P_N^2} \tag{9}$$

where P_N is the pressure of uncorrelated sources and N is the number sources. The combination of the uncorrelated sources does not depend on the phase of the pressure waves. Unlike the correlated soundwaves which are phase dependent for the output, uncorrelated waves will always give an increase in magnitude regardless of the phase.

2.5.7 Interaction of Soundwave with Objects

When soundwaves move though a medium, they interact with objects in their path. Depending on the properties of the material that it interacts with, the soundwave reflects, refracts, diffracts or get absorbed by the surfaces.

2.5.8 Absorption of Soundwaves in Air

Assuming a point source the wavefront that radiates and travels out in the form of an expanding sphere can be regarded as a spherical wavefront. In an ideal condition, the energy of this wavefront will be constant if there are no losses due to absorption by the medium. The energy of a soundwave is measured as the rate of energy transfer with respect to the area as expressed in (6). Since the surface area of the soundwave increases as the wavefront moves away from the source, the sound intensity reduces. In addition to the inverse square law, the energy of a soundwave is lost due to the combined action of the viscosity, heat conduction of air and the relaxation behaviour in rotational energy states of the molecules of air [20]. In addition to these factors, energy is lost due to humidity of the air. The attenuation of energy of a soundwave due to the effects of humidity and relaxation of behaviour in rotational energy states of the molecules are dependent on the frequency of the soundwave, and are known as excess attenuation [20]. The net absorption of sound energy in air is equal to the sum of the losses due to inverse square of (6) and the excess attenuation.

2.5.9 Reflection of Soundwaves

When soundwaves come in contact with an object that is larger than one fourth of the wavelength of the wave $\left(\text{greater than } \frac{1}{4}\lambda\right)$, the wave will be reflected [19, 20]. The reflection of the soundwave obeys the laws of reflection for any electromagnetic radiation. That is, when the wave bounces back from a smooth surface the angle of incident will be equal to the angle of reflection. When a soundwave is reflected from a surface the phase of the velocity components is changed. In addition to the wavelength of the wave, the other factor that affects the reflection of the soundwave is the rigidness of the material and the surface area of the material. As an example, materials that have a larger surface area like fibrous materials used in insulation of the walls and those materials that have holes like sponges or gypsum boards used in construction tend to have larger surface areas, and hence they absorb more energy from the wave and reflect less. When soundwaves come in contact with a surface that vibrates, part of the energy is lost due to frictional forces of the vibrating molecules within the material.

The concept of increasing the surface area for the soundwaves to interact has been used in the construction of flat walled Anechoic Chambers [22]. The walls of the chamber are covered with different density insulation materials layered such that the less denser material are at the outer most layers and the more denser materials are in the inner most layers. The different density materials absorb different frequencies of soundwaves. The amount of absorption by a material is given by the absorption coefficient 'a' expressed as [20]:

$$a = \frac{E_A}{E_I} \tag{10}$$

where E_A is the absorbed acoustic energy and E_I is the total incident acoustic energy. The value of 'a' is between 0 and 1, where 0 means all sound is reflected and 1 means all the sound is absorbed. This type of anechoic chamber is used in the experimental work of this thesis.

2.5.10 Refraction and Diffraction of Soundwaves

When soundwaves come in contact with objects that are one quarter of the wavelength or slightly less, the waves diffract around the object. That is the soundwave bends around the object. This bending of the waves is known as diffraction. Diffraction occurs due to variations in air pressure due to the inability of compressions and rarefactions in the soundwave to go to zero instantly after passing the edge of an object [20], causing part of the wave to continue to propagate and the wave to bend around the edges.

When soundwaves pass from one medium to another at an angle, the velocity of the soundwave changes at the boundary of the two mediums, this change in velocity occurs if the density or the temperature of the two mediums is different. This change in velocity of the soundwave causes it to change the direction of propagation according to Snell’s law. This change in direction of propagation is known as refraction of sound.

2.6 Soundfields

A soundfield is the space in which a soundwave propagates. There are several types of soundfields, these include:

- **Free Field**: A free field is uniform, where there are no boundaries and is free of other sound sources. In a free field, the sound energy flows in only one direction. In practice there are no ideal free fields naturally, but outdoor free spaces are considered free field. An anechoic chamber is a free field, since there is no reflection from any walls and there are no other sound sources.
- **Semi Reverberant Soundfields**: The concept of reverberation is based on the amount of reflections from the surroundings. In rooms where the walls and the furniture reflect and absorb, portions of the soundwaves may be considered a semi reverberant soundfield.
- **Reverberant Soundfields**: A room that has walls that are highly reflective and when there is very little absorption of the soundwaves in the room is considered a reverberant room. In a reverberant soundfield the time average of the mean square sound pressure is the same everywhere and the flow of energy in all direction is equally probable[20]. A person in a reverberant room first hears the original source without any reflections, known as the direct sound. The reflections that reach the person after the direct component is known as the reflected sounds, the number of times the reflections occur and delay between the reflections contribute to the amount of reverberation.

2.6.1 Direct Sound

When a source and a receiver are placed at opposite ends of a room, the sound that arrives at the receiver first is known as the direct path component. The path taken by the soundwave will be the shortest distance between the source and the receiver. The direct sound contains the actual information from the source without any contamination and is considered as sound in the free field. Since the direct sound is considered as free space it can be expressed according to (6), hence the intensity of the direct sound will be attenuated with distance according to the inverse square law. As a result, if the distance between the source and receiver is large than the direct sound component may be very small and interference by reflections can corrupt the direct component.

2.6.2 Early Reflections

The soundwave that bounces off walls and other objects in the room and reaches the receiver immediately after the direct sound is known as the early reflections. The early reflections cause interference and reduce intelligibility of speech. If the delay of the early reflection is more that 30ms then these are perceived as an echo [20]. The amount of early reflections depends on the surfaces of the room and the distance between the receiver and the surfaces. The early reflection, like the direct sound, behaves according to the inverse square law, in addition to the absorption effects of the surfaces and depending on the position of the receiver the intensity of the early reflections can vary.

2.6.3 Reverberant Sound

The sound that arrives after several reflections from all directions is known as the reverberant sound. These waves have been reflected off walls several times before they arrive at the receiver. The amount of reverberation depends on the distance between the source and receiver, the type of material used in the walls of the room and the size of the room. The time taken for the reverberations to die off is known as the reverberation time. Reverberation time RT_{60} is defined as the time taken for the sound energy to drop by 60dB compared to the direct sound and is expressed as:

$$RT_{60} = \frac{-0.161V}{S\ln(1-\alpha)} \qquad (11)$$

where S is the surface area, V is the volume, α is the abortion coefficient (typically $\alpha < 1$). Unlike the direct sound and the early reflection, the reverberant part of the sound remains constant, that is at any position in the room the intensity of the reverberant part will be the same. At any point in the room the receiver will receive reverberant sound from all directions and as a result there are a large number of soundwaves arriving at that point and their intensities are added together.

2.7 Sensors for Capturing Soundwaves

Soundwaves can be captured using sensors that can sense changes that occur in the propagation of the soundwaves. As described before, when a soundwave propagates there are changes in pressure, temperature and the density. Any sensor that can detect changes in pressure, temperature or the density can be used to capture a soundwave. The most common types of sensors that are used for capturing soundwaves are the:

- Temperature Sensors
- Pressure sensors

2.7.1 Temperature Sensors: Hot Wire Anemometer - Particle Velocity Sensors

The temperature sensors (Particle Velocity Sensor) such as the Microflown described in [23, 24] consist of two closely spaced silicon nitrate coated platinum wires which are heated to 300^0C. The separation between the wires is approximately 40 μm and the temperature difference of both wires is linearly dependent on the particle velocity [23]. The arrangement of the closely spaced hotwires is known as an anemometer. An anemometer senses the changes in temperature of the heated wires due to the passing soundwave. When a soundwave perpendicular to the heated wires passes over the wires, the wire that comes in contact with the soundwave first cools compared to the second wire. This change in temperature causes a change in resistance in the wires, which varies an output signal which is proportional to the particle velocity. The problem with a hot wire anemometer is that it cannot distinguish between two waves moving over it in opposite directions [25]. To overcome this problem, a steady bias mean air velocity is needed to give a signal that represents the particle velocity [25]. The disadvantage of this bias is that more noise is introduced to the output hence increasing the Signal to Noise Ratio (SNR). This increased noise in the output of a particle velocity sensor is one of the limitations for use in speech and other communication applications [25]. Furthermore, a particle velocity sensor is more sensitive to unsteady air flow compared to a pressure microphone. The AVS designed by Microflown is known as a P-U probe. Although the P-U probe can be used for source localization, the use of the P-U probes for capturing speech signals has not been documented [25].

2.7.2 Pressure Sensors

A microphone is a device that converts acoustical energy to electrical energy. Microphones are used in many different applications including capturing voice for communication and entertainment, in sonar and to detect seismic activity [26]. Microphones can be classified according to either directional characteristics or the mechanism used to convert the sound energy into electrical energy. For signal processing, the classification based on directional characteristics and the frequency response is much more useful that the mechanism used for sound conversion. Microphones can also be classified as pressure microphones, which respond to sound pressure with no regard to the direction of the soundwave, and the pressure gradient microphone which responds to both sound pressure and the direction of the soundwave. The frequency response of the microphone describes the voltage output of the microphone in decibels (dB) for different frequencies. For an ideal microphone, the frequency response is flat, over all frequencies.

2.7.3 Pressure Microphones

The ideal pressure microphones respond to the sound pressure with no effect on the output by the direction of the soundwave. When the diaphragm of a microphone is only exposed to a soundwave from one side the driving force on the diaphragm is given as:

$$F(t) = P_{ind}S \tag{12}$$

where P_{ind} is the directionless pressure, and S is the surface area of the diaphragm. From (12) it can be seen that there is no effect on the driving force from the direction of the soundwave.

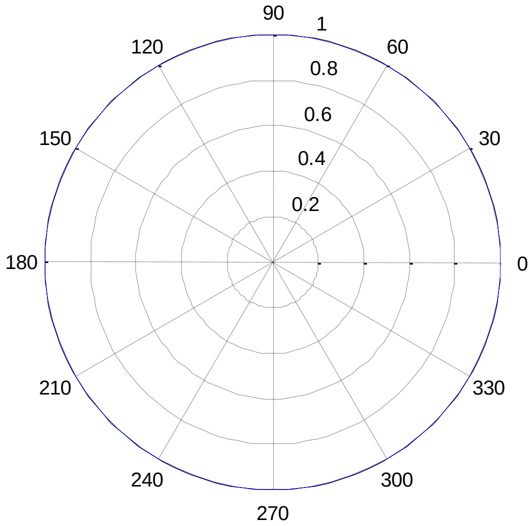


Figure 1: Polar plot of the response of an omni-directional microphone.

Pressure microphones contain a single opening and a single diaphragm which vibrate to vary either capacitance, resistance or the magnetic field in order to generate a time varying electrical signal. The most common pressure microphone is the capacitor microphone. Other types of pressure microphones include electret capacitor, dynamic microphones, and piezoelectric microphones. Figure 1 shows the polar plot of the response of an omni-directional microphone. From which it can be seen that the microphone captures the sound signals equally from all directions.

2.7.4 First Order Directional Microphones by Combining Pressure Microphones

First order microphones refer to any microphone that has a polar response equation that has a cosine term to the first power. In comparison, the second order microphone has a square of the cosine term. The first order microphone has a response proportional to the pressure gradient, whereas second order microphones have response that is proportional to the gradient of the gradient [26].

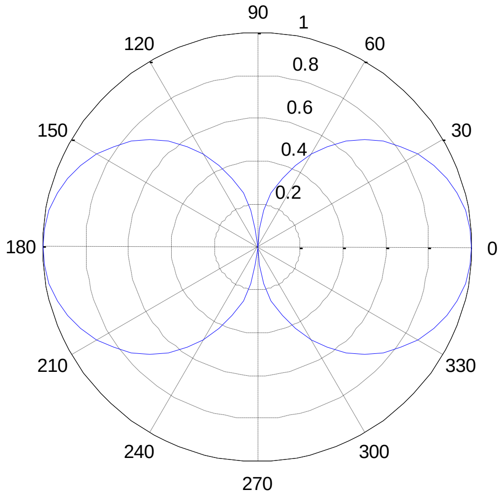


Figure 2: Polar plot of the response of a pressure gradient microphone.

A microphone can be formed by combining separate pressure and pressure gradient capsules separated with the diaphragm of the two elements aligned. This arrangement enables the control of the directional characteristics of the microphone response. The output from the system is from the linear addition of the two individual microphones. The root mean square of the output voltage is given as [26]:

$$E_{out} = \alpha(\beta + \gamma \cos\theta) \tag{13}$$

where α is a dimensional constant, β is the omni-directional component and γ is the pressure gradient component and θ represents the direction of the soundwave. By varying the values of β and γ, different polar patterns or directional characteristics can be achieved. The polar response curve can be obtained from the following equation [26]:

$$r = |\beta + \gamma \cos\theta| \tag{14}$$

where r is the radial distance from the origin between 0 and 1.The figure-of-eight pattern shown in Figure 2 known as the bidirectional pattern, is formed when $\beta = 1$ and $\gamma = 1$. As the contribution of β increases, the secondary lobe becomes smaller and smaller and at $\gamma = 0$ the polar response becomes an omni-directional microphone. For a first order gradient microphone there are two very important measures which are the Random Efficiency (RE) and the Distance Factor (DF). The RE is the measure of the on axis directivity in comparison to sounds arriving from all other directions. The DF is the measure of the reach of the microphone in a reverberant environment, relative to an

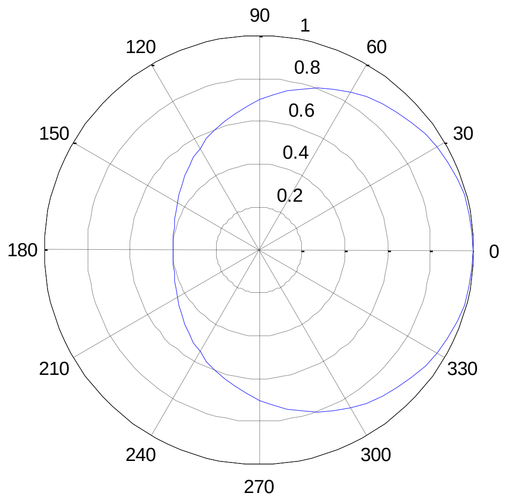


Figure 3: Polar plot of the response of a subcardioid microphone.

omni-directional microphone. The following are some ratios of β and γ and their directional characteristics:

- ***The subcardioid***: the values of β and γ are 0.7 and 0.3, and the directional response is directed to one side. These microphones are also known as a forward oriented Omni-directional microphone. The polar plot of a subcardioid is shown in Figure 3.
- ***The Cardioid:*** Shown in Figure 4 is formed by substituting the values of β and γ, as 0.5 and 0.5 respectively, in (14). The polar pattern is more forward focused and captures most of the sound from the forward direction while rejecting most sounds from the back. The cardioid microphone is the most commonly used microphone to capture speech and in musical performance.
- ***The Supercardioid:*** the values of β and γ are 0.37 and 0.63 respectively, in (14), this microphone captures from the front only and the front beam is narrower than that of a cardioid. The directional characteristics of the supercardioid microphone reduces the amount of reverberation captured and increases the strength of the on axis signal. The polar pattern of the response of a supercardioid microphone is shown in Figure 5.

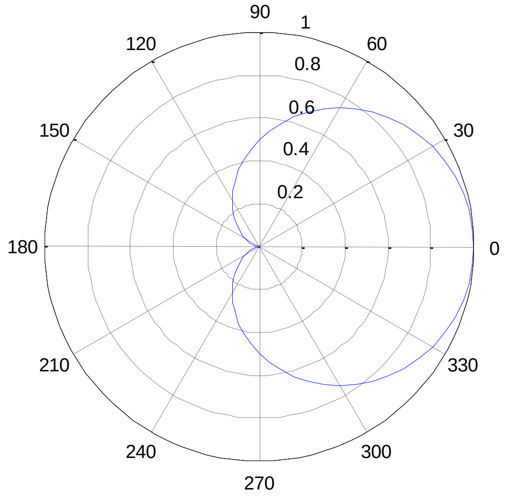


Figure 4: Polar plot of the response of a cardioid microphone.

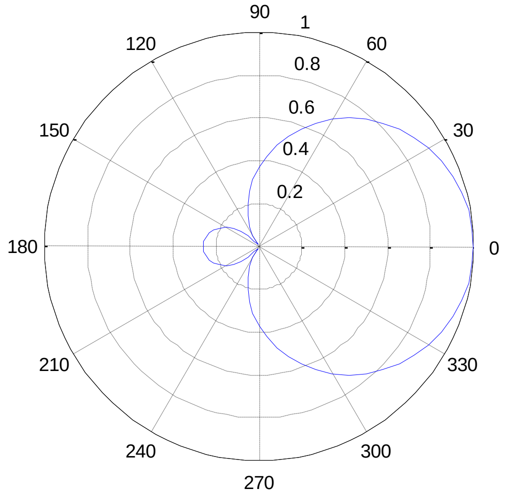


Figure 5: Polar plot of the response of a super cardioid microphone.

- ***The hyper cardioid:*** the values of β and γ are 0.25 and 0.75 respectively; this microphone captures the maximum from the forward direction, and provides the greatest rejection in a reverberant field. The polar pattern of the hypercardioid microphone response is shown in Figure 6.

For a hypercardioid, microphone the RE is ¼ which means that power distributed uniformly over all possible directions is ¼ that of the power captured from the on axis signal. For a hypercardioid the value of DF is 2 meaning that the working distance for a no axis signal is twice that of other directions.

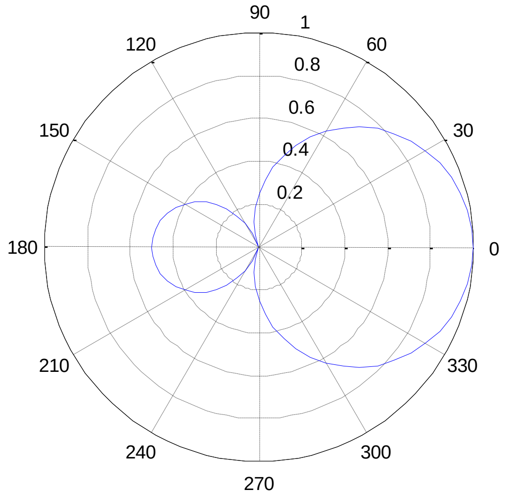


Figure 6: Polar plot of the response of a hyper cardioid microphone.

2.7.5 The Directional Characteristics of a Pressure Gradient Microphone

The pressure gradient microphone responds to acoustical pressure as well as the direction of the soundwave. Pressure gradient microphones are also known as velocity microphones. These microphones have openings on two sides of the diaphragm and sense the difference or gradient between the pressures on both sides of the diaphragm. The pressure difference between the two sides of the diaphragm is proportional to the velocity of the air particles of the soundwave. For a plane wave arriving at a gradient microphone, both sides of the diaphragm are exposed to the plane soundwave. In this case the diaphragm will capture the difference of pressure between the two sides of the diaphragm. The driving force on the diaphragm depends on the spatial rate of change of pressure rather than the pressure [27]. Figure 7 shows the diaphragm of a microphone exposed to a soundwave arriving at an angle θ.

When θ is $\frac{\pi}{2}$, the pressure on both side of the diagram will be equal, hence the driving force will be 0, and when θ is 0 or π the pressure on one side of the diaphragm will be maximum and the driving force will be equal to the surface area of the diaphragm multiplied by the pressure. The soundwave reaches the surface of the diaphragm that is not directly exposed by travelling around the diaphragm, and during

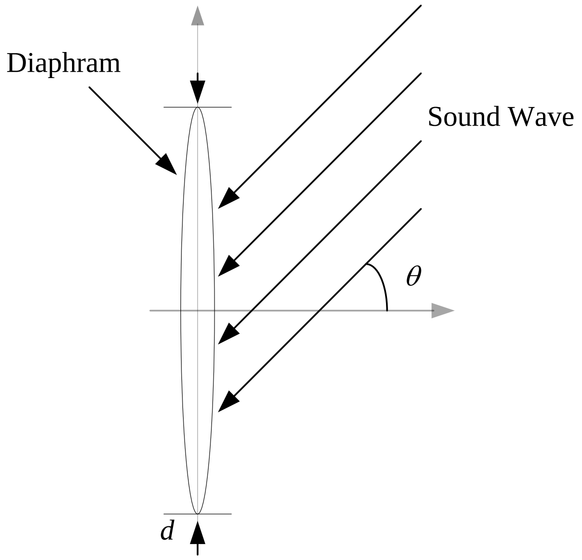


Figure 7: Diaphragm of a pressure gradient microphone.

this time the pressure of the soundwave changes. Hence, the driving force on the diaphragm will be the difference in pressure on both sides of the diaphragm multiplied by the surface area of the diaphragm. The pressure difference is the product of the spatial rate of change of acoustic pressure (pressure gradient) by the effective acoustic distance which is expressed as [27]:

$$P_{diff} = \nabla\, p \times d_{sep} \cos\theta = \frac{\partial}{\partial x} P_{dir} \times d_{sep} \cos\theta \tag{15}$$

where ∇p is the pressure gradient and d_{sep} is the acoustic distance separating the two sides of the diaphragm, the minimum being the diameter d of the diaphragm . The total driving force on the diaphragm is expressed as:

$$F(t) = \left[P_{dir} - \left(P_{dir} + \frac{\partial}{\partial x} P_{dir} d_{sep} \cos\theta\right)\right] S \tag{16}$$

$$F(t) = \left[-\frac{\partial}{\partial x} P_{dir} d_{sep} \cos\theta\right] S \tag{17}$$

where P_{dir} directional sound pressure and S is the surface area of the diaphragm. By substituting (4) into (17), the relation between the driving force and the particle velocity for a microphone with both sides of the diaphragm exposed to the sound pressure is given as:

$$F(t) = \left[-j v_{par} \rho_0 \omega d_{sep} \cos\theta\right] S \tag{18}$$

The above expression shows that the driving force on the diaphragm of a pressure gradient microphone is dependent on the acoustic particle velocity of the soundwave. If the soundwave on the diaphragm is from a source close to the microphone, then waves arriving on the microphone have a radial wavefront and can be expressed as:

$$P_{rad} = \frac{A}{r}e^{j(\omega t-kr)} \tag{19}$$

where A is a constant determined by the sound source and r is the radius of the wavefront. The ratio $\frac{A}{r}$ is the pressure amplitude which is dependent on the distance from the source. By substituting (19) into (17), the driving force exerted on the diaphragm of the microphone by a radial wavefront can be expressed as [27]:

$$F(t) = \left[-\frac{\partial}{\partial r}\frac{A}{r}e^{j(\omega t-kr)}d_{sep}\cos\theta\right]S \tag{20}$$

The solution for the differential part in (20) is [27]:

$$\frac{\partial}{\partial r}\frac{A}{r}e^{j(\omega t-kr)} = -\left(\frac{1}{r}+jk\right)e^{j(\omega t-kr)} \tag{21}$$

$$\frac{\partial}{\partial r}\frac{A}{r}e^{j(\omega t-kr)} = -\left(\frac{1}{r}+jk\right)P_{rad} \tag{22}$$

by substituting (22) into (20) the driving force on the diaphragm is given as:

$$F(t) = \left[\left(\frac{1}{r}+jk\right)P_{rad}d_{sep}\cos\theta\right]S \tag{23}$$

where $k = \frac{\omega}{c}$ is propagation constant, with $\omega = 2\pi$ and c is the phase velocity. The acoustic impedance of a plane soundwave is given as the ratio of the acoustic pressure to the particle velocity. The acoustic impedance is also equal to the density of air multiplied by the phase velocity of sound [27]. In the case of a radial wavefront the ratio of the acoustic pressure to particle velocity is given as:

$$\frac{P_{rad}}{v_{par(r)}} = \frac{j\omega\rho}{\frac{1}{r}+jk} \tag{24}$$

By substituting (24) into (23), the driving force on the diaphragm by a source close to the microphone is given as:

$$F(t) = \left[-jv_{par(r)}\rho_0\omega d_{sep}\cos\theta\right]S \tag{25}$$

Equation (25) shows that the driving force exerted on the diaphragm of a gradient microphone is independent of the type of the wave. That is whether the source is close to the microphone in the case of the radial wavefront or if the source is far in the case of a plane wave.

2.7.6 The Proximity Effect of a Pressure Gradient Microphone

In Section 2.7.3 it has been shown that the driving force exerted on the diaphragm is independent of the source to microphone distance. Even though the driving force is independent of the source to microphone distance there is a phenomenon known as the proximity effect for the gradient microphone which is the relation between the particle velocity, frequency of the soundwave and the separation of the source and microphone. By rearranging (24) the particle velocity is expressed in terms of pressure and the other terms as shown:

$$v_{par(r)} = \frac{1+jkr}{jkr} \times \frac{P_{rad}}{\rho c} \tag{26}$$

By taking the magnitude of the particle velocity and considering the separation between the source and microphone to be large or the wavelength of the soundwave is small, then (26) will reduces to [27]:

$$\left|v_{par(r)}\right| = \frac{\sqrt{1+k^2r^2}}{kr} \times \frac{|P_{rad}|}{\rho c} \tag{27}$$

$$\left|v_{par(r)}\right| = \frac{|P_{rad}|}{\rho c} \tag{28}$$

From (28) it can be seen that the particle velocity is directly proportional to the acoustic pressure. For the case where the separation between the microphone and the source is small or the wavelength of the soundwave is large (26) becomes [27]:

$$\left|v_{par(r)}\right| = \frac{1}{\omega r}\frac{|P_{rad}|}{\rho c} \tag{29}$$

In this case, the particle velocity is inversely proportional to the frequency, meaning when the source is close to the microphone, lower frequencies will produce larger responses.

2.8 The Microphone Array

A microphone array consists of multiple microphones arranged in a pattern to form a desired polar response, or to get a combined output from all the microphones. There are several geometrical patterns used in microphone arrays, like the Uniform Linear Array (ULA) which is the most common and widely used microphone array. Other microphone arrays include circular microphone arrays, spherical microphone arrays and Soundfield microphone arrays. Microphone arrays can be divided into two categories, which are:

a) Distributed arrays where microphone capsules are geometrically distributed.

b) Co-located microphone arrays where the microphone capsules are arranged such that there is no delay between the sounds reaching the capsules in the array.

Traditionally, microphone arrays were primarily used for detecting the DOA estimates for sound sources. In recent studies it has been found that the multichannel nature of the microphone arrays can be successfully used for signal enhancement, noise removal and source separation [28, 29].

There are some terms that are important for microphone arrays, these include the array aperture, beam pattern or directivity pattern, beam width and array gain which are explained below [30].

- ***Array Aperture:*** is the spatial region around an array that receives the soundwaves. The term originates from antenna theory where the term is referred to the spatial region that transmits or receives the signals. In antenna theory, a transmitting aperture is termed an *active aperture* and receiving aperture is termed a *passive aperture*.
- ***Beam Pattern or Directivity Pattern:*** the main aim of forming an array is to create a system that is directional. Hence any array can be said to be directional in nature; this is because the amount of received or transmitted signal from the arrays varies with the direction. The directivity of an array is a function of frequency and directivity.
- ***Beam Width:*** is the angle between the half power points or the -3dB point of the main lobe. This definition is the standard definition used in antenna theory and the same definition is used in microphone arrays.
- ***Array Gain:*** The improvement to the Signal to Noise Ratio (SNR) between a reference sensor and the array output.

2.8.1 The Concept of Near and Far Field

The distance between the source and array and the length of the array is an important factor in the derivation of many DOA estimation algorithms; this assumption is known as the far field assumption. With regards to the microphone array, the far field assumption is where if the source to the array separation is much larger than the array

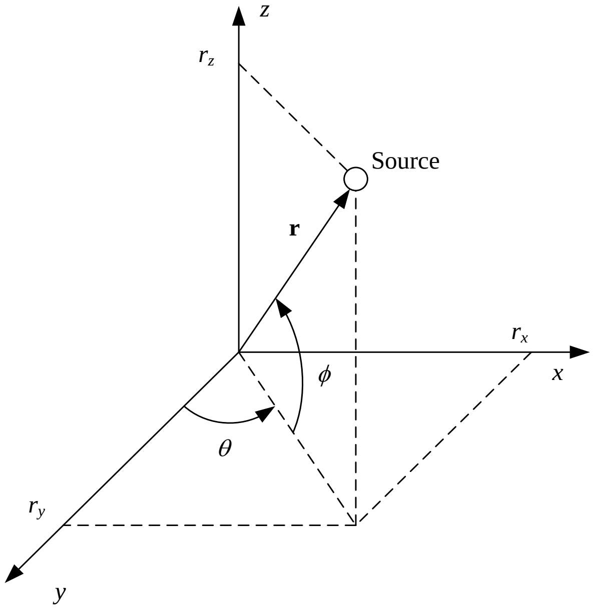


Figure 8: 3D representation of a source location.

dimensions, then the source can be assumed to be in the far field [31]. The far field assumption can be mathematically expressed as:

$$r \gg \frac{2D}{\lambda}, \tag{30}$$

where r is the distance between the source and the array and D is the length of the array and λ is the wavelength.

The assumption made here is the curvature of the wave arriving at the array is small compared to the array, hence the wavefronts are planar. In the case where the source and array are close, such that the separation is comparable to the array size, the curvature of the wave is significant, and hence the source will be assumed to be in near field.

2.8.2 Signals in Three Dimensional Space

The position of a source in three dimensional space can be expressed in polar or Cartesian coordinates. Figure 8 shows the position of the source relative to the array. The vector **r** can be expressed in terms of the azimuth and elevation angles as [32]:

$$r_x = ||\mathbf{r}||sin(\emptyset)cos(\theta) \tag{31}$$

$$r_y = ||\mathbf{r}||cos(\emptyset)cos(\theta) \tag{32}$$

$$r_z = ||\mathbf{r}||sin(\emptyset) \tag{33}$$

where r_x, r_y and r_z are position of the source and θ and ϕ are the azimuth and elevation angles respectively.

2.8.3 The Uniform Linear Array (ULA)

The most common microphone array is the Uniform Linear Array (ULA). The omni-directional microphone capsules in a ULA are arranged in a straight line with a separation of d between the capsules, as shown in Figure 9. The far field directivity pattern of a ULA is expressed according to [26, 28]:

$$BP = \left|\frac{\sin[N\pi f d(\cos\varphi - \cos\theta)/c]}{N\sin[\pi f d(\cos\varphi - \cos\theta)/c]}\right| \quad (34)$$

where N is the number of capsules in the array, d is the separation between the capsules, c is the speed of sound in air and f is the frequency of the incident wave. Parameter φ is $0 \leq \varphi \leq \pi$ and θ is the direction of arrival. From (34) it can be seen that the directivity pattern of a ULA is a function of frequency as well as the separation between the capsules. The useful range of a ULA array is up to $\frac{d}{\lambda} = 1$, and after this the array starts to exhibit off axis lobes. To overcome this problem of off-axis lobes, the capsules are spaced logarithmically such that the capsules are closer together towards the centre as proposed by Van der Wal et al [33]. Figure 10 shows the polar plot of the beam pattern generated for four sensors separated by 3 cm at a frequency of 2 kHz.

Figure 10 shows the beam width is wider and hence interferences around 60 degrees and 120 degrees will be picked up by the array. From (34) it can be seen that by increasing the separation or increasing the number of elements or both in the array, a sharper beam width can be obtained. A sharper beam means that the array will be able focus more on the source and minimise the interference.

Figure 11 shows the effect on the beam pattern of the array when separation is increased from 3 cm to 12 cm and by increasing the number of microphones from 4 to 12 microphones. From Figure 10 and Figure 11 it is seen that to achieve higher directionality, the array length has to be increased or the number of elements has to be increased. As mentioned before, as the number of the elements or the size of the array is increased the off axis lobes start appearing. These off axis-lobes will capture sources that are in the direction of the off-axis lobes, which will introduce unwanted noise and reverberations.

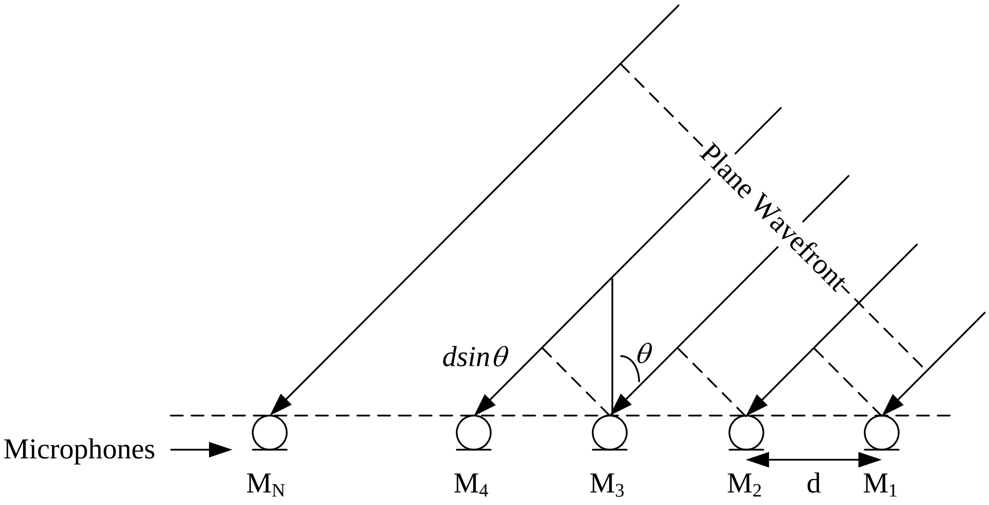


Figure 9: ULA microphone array.

2.8.4 The Output of an ULA

The signals from a ULA can be expressed in terms of the signal arriving at the m^{th} microphone as follows:

$$x_m(t) = h_m * s_i(t - \tau_m) + n_m(t) \qquad (35)$$

where $x_m(t)$ is the signal arriving at the m^{th} microphone, h_m is the factor representing the impulse response from source to the m^{th} microphone, $s_i(t - \tau)$ is the delayed version of the i^{th} source signal compared to the reference microphone. In most cases the reference microphone is the 1st microphone and $n_m(t)$ is the noise at the m^{th} microphone. Here it is assumed that there are L sources and $L \leq M$, where M is the number of microphones in the array. In the frequency domain the matrix notation of the signals arriving at the microphone can be expressed as:

$$\boldsymbol{X(\omega)} = \boldsymbol{A(\omega)} * \boldsymbol{S(\omega)} + \boldsymbol{N(\omega)} \qquad (36)$$

The vector $\boldsymbol{A}$ is known as the steering vector which is expressed as:

$$A(\omega) = [\, e^{j\omega d_1 (\sin\theta)/c} \; e^{j\omega d_2 (\sin\theta)/c} \;\ldots\ldots e^{j\omega d_m (\sin\theta)/c}] \qquad (37)$$

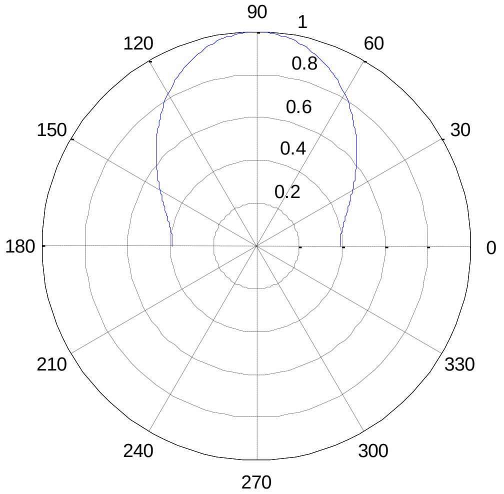


Figure 10: Polar Plot of the beam pattern for a ULA with 4 sensors separated by 3cm

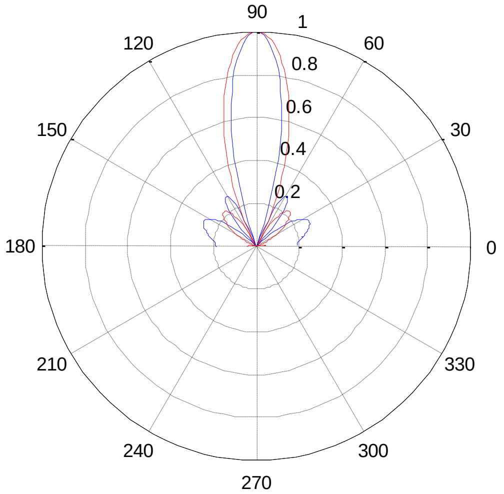


Figure 11: Polar Plot of the beam pattern for a ULA with 12 sensors (red), with separation of 12cm (blue).

where ω is the time frequency and d_m is the distance between the reference microphone and the m^{th} microphone, c is the speed of sound in air and angle θ is the DOA estimate in azimuth. The $\sin\theta$ term in (37) is known as the Time Difference of Arrival (TDOA) τ, expressed in seconds where:

$$\tau = \frac{d_0\,(\sin\theta)}{c} \tag{38}$$

2.8.5 Circular Microphone Array

The microphones in a circular microphone array are arranged in the circumference of a circular structure, which can be a solid structure or a frame. The difference between the two structures depends on the application. The solid structure is often used for capturing two dimensional signals while the circular frame is used for DOA estimation applications. A circular microphone array generally contains a larger number of microphones compared to a ULA. The advantage of the circular structure is it can accommodate the large number of microphone capsules in a small space. As an example the circular array described in [34] and [35] contain 32 to 288 microphones in 0.5m and 1m diameter arrays. If similar numbers of microphones with similar separation were to be used in a ULA, the array sizes will be 3.14m and 6.14m. Some applications of the circular array other than DOA estimation [34] [36] include panoramic [37] and ambisonic recording [35] of the soundfield. The type of capsules used in the construction of the circular microphone array depends on the application. In [34], omni-directional capsules are used and in [35] uni-directional (cardioid family) microphones are used.

2.8.6 Spherical Microphone Array

The spherical microphone arrays are generally used for three dimensional recordings due to its three dimensional symmetry [38]. The aim of the spherical array is to capture the sound information in the three dimensional space as accurately as possible. In addition to capturing three dimensional sounds spherical microphone arrays can be used for beamforming for speech enhancement and DOA estimation. The advantages of the spherical arrays is it can house a large number of microphones in a small space compared to any other microphone array and can be used to steer beams to any directions [39] in three dimensional space. Most spherical arrays are constructed using either omni-direction microphones or cardioid microphones. The microphone positions can either be random as in [40] or it can be positioned to get the best performance as in [38].

2.8.7 Soundfield Microphone Array

The Soundfield microphone array contains four cardioid microphones arranged in a tetrahedron configuration. The soundfield microphone has four outputs which are the X, Y, Z and W components which are also known as the B format output. The B format outputs are formed by combining the outputs from the four cardioid microphones. The main use of the soundfield microphone is to record three dimensional surround sound. The difference between the soundfield and the circular or the spherical array is that the soundfield is able to record a three dimensional soundfield at studio quality using only four microphones, whereas the circular and spherical arrays use a large number of microphone capsules and the size of array is large compared to that of the soundfield.

The four microphones that form the Soundfield microphone array are named as Front Left (FL), Front Right (FR), Back Left (BL) and Back Right (BR). The Left Front microphone and the Right Back are back to back but tilted symmetrically from vertical. Similarly the Right Front and the Left Back microphones are back to back tilted downwards.

A figure-of-eight response can be formed in the horizontal plane with axis along the LF and RB line by subtracting the outputs from the LF and RB. Similarly the RF and LB can be subtracted to form a horizontal figure-of-eight response with the axis in the line along the RF and LB line. By adding the two figure-of-eight patterns and LF and RF in phase the X component of the B format output can be formed. Similarly the other B format signals from the output of the microphone capsules can be formed using (39) to (42).

$$\mathrm{W} = \mathrm{LF} + \mathrm{RB} + \mathrm{RF} + \mathrm{LB} \tag{39}$$

$$\mathrm{X} = \mathrm{LF} - \mathrm{RB} + \mathrm{RF} - \mathrm{LB} \tag{40}$$

$$\mathrm{Y} = \mathrm{LF} - \mathrm{RB} - \mathrm{RF} + \mathrm{LB} \tag{41}$$

$$\mathrm{Z} = \mathrm{LF} + \mathrm{RB} - \mathrm{RF} - \mathrm{LB} \tag{42}$$

2.8.8 Co-located Microphone Arrays

The microphone arrays that have been discussed so far have the capsules spatially distributed. A co-located microphone array such as the AVS array that is the topic of study in this thesis, has its microphone capsules arranged such that the wavefront arrives at all the microphone at the same instance in time. These microphone arrays generally are extremely compact; contain directional microphone capsules, and are generally used for source localization. Detail analysis of co-located microphone array will be presented throughout this thesis.

2.9 Microphone Array Signal Processing

For the microphone array shown in Figure 9, the source is assumed to be in the far field and soundwaves arrive at an angle θ perpendicular to the array, where the soundwave reaches the microphone M_1 first and after a small delay it arrives at M_2. The delay is the time taken for the soundwave to travel a distance $d \sin\theta$. This delay is known as the Time Difference of Arrival (TDOA) τ as expressed in (38). The TDOA is one of the most important parameter that can be extracted from a spatially distributed microphone array, it enables estimation of the DOA and is critical in most beamforming algorithms. There are two ways in which τ can be calculated,

1. The delay between each pair of the microphones
2. The delay between the reference microphone and the m^{th} microphone

The latter is used in most applications as the accuracy of the TDOA estimate increases with the increase in separation. Furthermore for TDOA to be useful the array geometry has to be known.

2.9.1 Direction of Arrival Estimation for a General Microphone Array

The DOA estimation of the source is the first step to many other speech enhancement algorithms like beamforming, dereverberation and blind source separation. These algorithms rely heavily on the accuracy of the DOA estimation stage for their performance. There are several DOA estimation algorithms and most of these algorithms are tailored to specific array geometries. There are very few universal algorithms that can be directly applied to all microphone arrays regardless of the array geometry. As an example, DOA estimation based on TDOA can be calculated in any array for pairs of microphone that are spatially separated. In general, the performance of the algorithm largely depends on the following factors [31]:

1. The number of microphone capsules in the array
2. The array configuration (the positions of the microphones and their separation)
3. Number of sources and types of sources
4. The characteristics of the room (amount of reverberation)
5. Amount of background noise (diffuse noise) from fans computers other similar sources
6. If the sources are stationary or mobile

To improve the accuracy of the DOA estimate in adverse conditions the number of microphones can be increased. But this is not always practical as conditions in a real environment can change rapidly and unexpectedly. The challenge is to build an array and an estimation technique that can be used in most condition with good accuracy.

The DOA estimation techniques can be broadly divided into three main techniques, which are:

1. TDOA based approaches
2. Steered Power Response approaches
3. Spectral Estimation approaches

2.9.2 TDOA Based Approaches

The approach of TDOA for source localization is based on two criteria which are:

1. There are pairs of microphones with the separation between them known.
2. The pairs are spatially distributed with their location relative to each other known.

The time delay estimate of the speech signal for each pair of microphone is calculated and using the location information of the pairs of microphones, the DOA estimate can be calculated. Hence, these approaches are only applicable to microphone arrays that are spatially distributed. One of the most interesting aspect of this technique is that no matter how the microphones are arranged, as long as the position information of the microphones is known, the DOA estimates can be calculated. Hence, this algorithm can be applied to any spatially distributed microphone array. The drawbacks of this technique are lower performance in the presence of considerable background noise and room reverberation [31]. The time delay estimates from the pairs of the microphones are calculated using the cross correlation function.

2.9.3 Cross Correlation

The arrangement of the microphones in spatially distributed arrays such as the ULA, circular and spherical microphone arrays is such that any two microphones will form a two element ULA array. Let the signals from microphone M_1 and M_2 from the ULA in Figure 9 be x_1 and x_2. The cross correlation between the two signals is expressed as:

$$x_1(t) * x_2(t) = \int_{-\infty}^{\infty} x_1^*(\vartheta)\, x_2(t-\vartheta)d\vartheta \quad (43)$$

where $*$ represents complex conjugate of $x_1(t)$, then the maximum of the cross correlation function of (43) will occur when the two signals are perfectly aligned, hence the TDOA τ can be expressed as:

$$\tau = \text{argmax} \int_{-\infty}^{\infty} x_1^*(\vartheta)\, x_2(t-\vartheta)d\vartheta \quad (44)$$

At low levels of diffuse noise and low levels of reverberation the DOA estimate calculated using this method is accurate but as the amount of noise and reverberation increases the accuracy of DOA estimates starts to suffer due to errors in calculating the TDOA from (44) [31]. The changes that can improve the accuracy of the DOA estimate are:

- Increasing the number of microphones in the array
- Increasing the separation of the microphone

Unfortunately these changes are not practical in real situations as the amount of noise and reverberation can change due to changes in the environment. The alternative to changes in the array design is the use of an improved cross correlation function by deemphasising the frequency dependent weightings as proposed in [31, 41, 42] known as the Phase Transform (PHAT) which reduces errors in noisy and reverberant conditions.

2.9.4 Generalized Cross Correlation with Phase Transform (GCC-PHAT)

The value of TDOA can be found by applying the modified version of the cross correlation function described in [31, 41, 42] known as the Generalized Cross

Correlation with Phase Transform (GCC-PHAT) . The advantage of using the GCC-PHAT algorithm is it offers more resistance to errors in noisy and reverberant conditions [43]. The GCC PHAT places equal emphasis on each component of the cross spectrum phase, the peak in the GCC PHAT spectrum corresponds to the dominant delay. For the microphone array in Figure 9 the cross correlation given by (44), the Generalized Cross Correlation (GCC) of x_1 and x_2 , $R(\tau)$, can be obtained from the cross correlation of the filtered versions of x_1 and x_2 as in [41]. Let the filters be h_1 and h_2 then the GCC function can be expressed in terms of Fourier transforms of x_1 and x_2 as:

$$R(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\big(H(\omega)X_1(\omega)\big)\left(H(\omega)X_2(\omega)\right)^{*}e^{j\omega\tau}d\omega, \tag{45}$$

$$R(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\psi X_1(\omega)X_2(\omega)^{*}e^{j\omega\tau}d\omega, \tag{46}$$

where $X_1(\omega)$ and $X_2(\omega)$ are the Fourier Transforms of the microphone outputs and $H_1(\omega)$ and $H_2(\omega)$ are Fourier transform of the filters, and $\psi = \frac{1}{|X_1(\omega)X_2(\omega)^{*}|}$ is the phase transform weighting. The TDOA is calculated as follows:

$$\tau = \ \text{argmax}\, R(\tau). \tag{47}$$

The described algorithm has been used to get accurate results for source localization in reverberation and in diffuse noise, but when there is significant amount of diffuse noise or reverberation and when there is more than one source the performance of this algorithm suffers. To improve performance in these conditions, several improvements have been proposed. A modified versions of the GCC PHAT implementation to improve performance of the DOA estimates in noisy conditions are presented in [44, 45] and in [46] a modified version of the GCC PHAT algorithm is used for DOA estimation of multiple sources.

2.9.5 Steered Power Response Approaches

The Steered Power Response (SRP) can be defined as combining all the signals from the array to get the Maximum Likelihood (ML) estimate such that the maximum signal energy from a given direction is obtained. The idea behind SRP is to beamform all pairs of microphones in the array and then combine the pairs together. The simplest method for beamforming is the Delay and Sum (DS) Beamformer. The beamformer provides some enhancement to the signal while the noise and reverberation components are attenuated to some extent. For microphones in the ULA array of Figure 9 the, DS beamformer can be expressed as:

$$S(t) = \sum_{m=1}^{M} x_m(t - \Delta t) \tag{48}$$

where Δt is the steering delays in relation to the reference microphone $(m = 1)$ and $S(t)$ is the beamformed output. The effectiveness of this beamforming operation is minimal as the beamformer is not able to enhance the target even in moderate levels of reverberation and noise. When this approach is applied to a co-located microphone array (such as an AVS) the beamforming approach simply becomes a summing operation, as the channels in the co-located microphone array are time aligned. More advanced versions of the beamformers are those that perform filtering before the summing of the channels and these beamformers are known as the filter and sum beamformers. The role of the filter in the filter and sum approach is to minimize the SNR in noisy conditions. Most beamformers that are available fall into this category and these beamformers can be applied to both co-located and spatially distributed microphone arrays. For the microphone array in Figure 9, the filter and sum beamformer can be represented as:

$$S(\omega) = \sum_{m=1}^{M} H_m(\omega) X_m(\omega) e^{j\omega\Delta t}, \tag{49}$$

where $H_m(\omega)$ and $X_m(\omega)$ are the Fourier transform of the filter and the m^{th} channel of the microphone array respectively. The DOA estimates from (49) is found by finding Δt that produced the maximum energy in the output of (50).

$$\tau = \operatorname{argmax}_{\Delta t} \int_{-\infty}^{\infty} |S(\omega)|^2. \tag{50}$$

An enhancement based on the filter and sum approach which is similar to that of the CGG PHAT algorithm is proposed in [47] which is known as the Steered Power Response Phase Transform (SRP PHAT). This is one of the most widely used algorithms for source location.

2.9.6 Steered Power Response Phase Transform (SRP PHAT)

The filter and sum approach proposed when applied to a pair of microphones in the array is exactly the same as that which has been presented in (45) and (46). Extending these equations to include all the microphone pairs in the array, the energy of the combined array can be expressed as:

$$E(\Delta t) = \sum_{l=1}^{M} \sum_{m=1}^{M} \int_{-\infty}^{\infty} \psi_{lm} X_l(\omega) X_m^*(\omega) e^{j\omega\Delta t} d\omega \tag{51}$$

from (51) it can be seen that the SRP PHAT is in fact GCC PHAT applied to individual pairs of the microphones which are then combined. The important point to be noted here is that no matter what the array geometry, as long as the microphones are spatially distributed the SRP PHAT can be used for DOA estimation. The SRP PHAT algorithm is one of the most robust DOA estimation algorithms used. It has shown good performance in noisy and reverberant environments. The most important assumption in the use of the SRP PHAT algorithm is the spatially distributed array with a large numbers of microphone capsules. Hence the SRP PHAT algorithm can only be used in the array that are large and with larger number of microphone capsules like those described in Section 2.8.4, 2.8.5 and 2.8.6 and due to these basic assumption the SRP PHAT algorithm cannot be used with co-located microphone arrays such as the AVS. There are several improvements that have been proposed to the SRP PHAT which include stochastic region contraction approaches proposed by [48-50] for multiple source location and improvements to the robustness in steering has been proposed by [51].

2.9.7 Spectral Estimation Based Approaches

The spectral estimation methods for DOA estimation can be used with any type of microphone array. These methods can generally be classified as shown in Figure 12 [52]. Unlike the TDOA based methods, spectral estimation methods can be applied to co-located microphone arrays like the AVS. The spectral estimation methods used in DOA estimation are based on Autoregressive Modelling (AR), Minimum Variance (MV) methods and Subspace methods such as the MUltiple SIgnal Classification (MUSIC) method. The basic concepts of all of these algorithms are to maximize the likelihood that a signal arrives from a given direction. The attractiveness of the maximum likelihood estimation is that the there is no restriction on the number of the sensors and sources; that is these algorithms theoretically can be used when the number of source are more than the number of microphones.

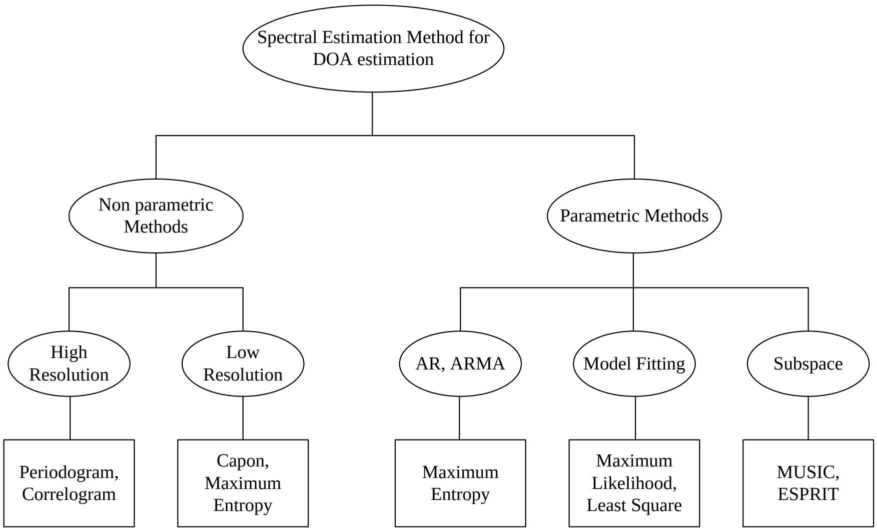


Figure 12: Classification of spectral estimation based DOA estimation algorithms

2.9.8 Maximum Likelihood Estimator

In this section, the Maximum Likelihood (ML) Estimator for a general microphone array will be derived. This derivation will lead to a possible DOA estimate from the output signals of the array. Let the microphone array output be represented as:

$$\mathrm{X(t) = A(\theta)S(t){+}N(t)} \tag{52}$$

where $\mathbf{X} = [x_1(t)\ x_2(t)\ \ldots x_M(t)]$ is the matrix of outputs from the microphones in the array with M microphones, $\mathbf{A}(\boldsymbol{\theta}) = [\boldsymbol{\alpha_1}\ \boldsymbol{\alpha_2} \ldots. \boldsymbol{\alpha_M}]$ is the general form of the steering vector which is $1 \times M$ long. $\mathbf{S} = [s_1\,(t)\ s_2(t) \ldots s_l(t)]$ represents the L sources that arrive at the array and it is assumed that $M \leq L$. $\mathbf{N} = [n_1\,(t) n_2(t) \ldots n_n(t)]$ is the noise matrix, here it is assumed that the noise is Gaussian and white with zero mean and variance is $\sigma^2$and N is the number of samples in each frame of data. The unknown for which the maximum likelihood estimator is found is θ, which is the possible location of the target source. The Probability Density Function (PDF) for (52) can be expressed as [53]:

$$pdf(\boldsymbol{X(t)}) = \sum_{t=1}^{N} \frac{1}{\pi det(\sigma^2 I)} e^{\left\{-\frac{1}{\sigma^2}|\boldsymbol{X(t)} - \boldsymbol{AS(t)}|^2\right\}} \tag{53}$$

where $|\boldsymbol{X} - \boldsymbol{AS}|^2$ is the noise power to be minimized and I is the identity matrix. If $R_n = E(NN^H) = \sigma^2 I$ and I is the identity matrix, R_n is the noise covariance matrix, the PDF function of (53) in terms of R_n can be expressed as:

$$pdf(X(t)) = \sum_{t=1}^{N} \frac{1}{\pi \det{(R_n)}} e^{(\boldsymbol{X(t)} - \boldsymbol{AS(t)})^H R_n (\boldsymbol{X(t)} - \boldsymbol{AS(t)})} \tag{54}$$

The normalized log likelihood function for (54) can be expressed as [54]:

$$L(\theta, R_n, S(t)) = -\log(\det(R_n)) - \sum_{t=1}^{N}(\boldsymbol{X(t)} - \boldsymbol{AS(t)})^H R_n(\boldsymbol{X(t)} - \boldsymbol{AS(t)}) \qquad (55)$$

By solving (54) with respect to each of the variables and maximizing the function, the most probable estimate of each of the variables can be found. A detailed description of the log maximizing function can be found in [54-57].

The advantages of using the algorithms derived from the ML estimator include accurate DOA estimation; DOA estimation for more than one source, can be applied to any array geometry. The only drawback of the algorithm is that it is computationally complex compared to other DOA estimation algorithms described previously. The two most commonly used DOA estimation algorithms based on the ML estimator are the MUSIC algorithm and the Estimation of Signal Parameters via Rotational Invariance Technique (ESPRIT) algorithm.

2.9.9 The MUSIC Algorithm for DOA Estimation

The MUSIC algorithm for DOA estimation is one of the most popular DOA estimation algorithms. The MUSIC algorithm splits the array output covariance function into the signal and noise components using Eigen-decomposition. For the outputs of the array $\boldsymbol{X(t)}$ the covariance matrix R is found under the assumption that source are uncorrelated [32, 58].

$$R = E[X(t)X(T)^*] \qquad (56)$$

$$R = E[AS(t)A^H S(t)] + E[N(t)N(t)^H] \qquad (57)$$

$$\text{Let } \hat{S} = E[S(t)S(t)^H] \qquad (58)$$

The rank of $\hat{S}$ is q, substituting $\hat{S}$ into (58) gives [58, 59]:

$$R = E\left[A\hat{S}A^H\right] + R_n \qquad (59)$$

The correlation matrix of the sources R_s from (59) can be defined as [58, 59]:

$$R_s = E\left[A\hat{S}A^H\right] \qquad (60)$$

where H stands for the Hermitian transpose. The Eigen decomposition of source covariance matrix R_s will result in set of Eigen values and Eigen vectors. Some of these Eigen values will be equal to zero, the Eigen vectors corresponding to these zero Eigen values are Q_0. The concept of the MUSIC algorithm is Eigen values of R_s that correspond to the Eigen values Q_0 are orthogonal to the M steering vectors of $\boldsymbol{A}$. Thus pseudo-spectrum of the MUSIC algorithm can be expressed as [59]:

$$P(\theta)_{MUSIC} = \frac{1}{\sum_{m=1}^{M}\left|Q_0^* A(\theta)\right|^2} \qquad (61)$$

Since Eigenvectors Q_0 are orthogonal to the steering vectors, when a source is found at θ the denominator of (61) approaches zero, hence a maxima occurs. The M largest peaks in (61) correspond to sources. In practice, the source covariance matrix R_s is not available hence the algorithm relies on the covariance matrix R of the array output. The Eigen decomposition of the covariance matrix R of the array output can be expressed as [32, 60-62]:

$$R = Q\Lambda Q^H \tag{62}$$

$$R = Q(\Lambda + \sigma^2 I)Q^H \tag{63}$$

The Eigenvector Q from (63) can be divided into Q_s which is the Eigenvectors due to source and Eigenvectors due to noise Q_n. This partitioning of the Eigen vector into two subspaces is the differentiating characteristic of subspace methods compared to other DOA estimation methods. The corresponding noise Eigen value will be σ^2, which corresponds to the smallest Eigen value and due to the orthogonality of Q_s and Q_n, the noise Eigen vectors are orthogonal to the steering vectors. Hence, by substituting the smallest Eigen values form the output covariance matrix into (63), the DOA estimates can be found.

Variants of the MUSIC algorithm are the ROOT MUSIC algorithm [63], spectral smoothing MUSIC [64, 65] and the cyclostationarity MUSIC [66]. The ROOT MUSIC is a model based algorithm, where in the case of DOA estimation the model is assumed to be the steering vector. The spectral smoothing MUSIC algorithms improve the performance of the MUSIC algorithm when the sources are correlated and cyclostationarity MUSIC enables improved performance with reduced array elements.

2.9.10 Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT)

The ESPRIT algorithm is a subspace based algorithm for DOA estimation. Unlike the MUSIC algorithm, the ESPRIT algorithm does require exhaustive search though all possible steering vectors to obtain the DOA estimate [67]. In addition the signal subspace is estimated from the data matrix rather than then the correlation matrix [32]. The ESPRIT algorithm is more complex than the MUSIC algorithm as the ESPRIT algorithm requires two Eigen Decompositions and relies heavily on matrix manipulation. Detailed derivation of the ESPRIT algorithm can be found in [59, 67].

The DOA estimation algorithms that have been discussed in this section were designed for antenna arrays and sonar applications. The modified versions of these algorithms have been applied to speech sources as in [68].

2.9.11 Beamforming

Beamforming can be defined as the process of combining the output signals from an array of sensors with a weighting function such that a source in a given direction is emphasised while other sources in other directions are attenuated. The expression given in (48) is the most general form of the beamformer, where the signals are delayed in time and added. There are several forms of beamformer that are designed for different types of sensor arrays, which will be discussed later in this section. In general, beamformers can be classified according to how the weights are obtained as either data independent or statistically optimum (data dependent) [69].

2.9.12 Data Independent Beamformers

The weights of the beamformer for the data independent beamformer are chosen such that the output of the beamformer is a close approximation of the desired signal. The weights have no relation to the actual data from the output of the array. The analogy of the data independent beamformer is Finite Impulse Response (FIR) filtering, where the filter is designed to extract the desired part of the signal. The DS beamformer or the filter and sum beamformer can be thought of as a data independent beamformer.

2.9.13 Statistically Optimum Beamformers

Statistically optimum beamformers are designed based on statistics of actual signal properties, location information, interferers and noise signals that are received at the array. The first statistically optimum beamformers are the reference signal based beamformer, Multiple Sidelobe Canceller (MSC) and the Maximization of the Signal to Noise Ratio Beamformer (MSNR). The beamformers described above require reference signal, interferences and noise signals which in practice are not available or not known. The two main approaches in statistically optimum beamformers that were proposed to overcome these shortcomings are Minimum Mean Square Error (MMSE) and Linearly Constrained Minimum Variance (LCMV) based beamformers.

The LCMV beamformer has been presented by many authors in different ways [9, 70-75]. The main goal of the beamformer is to constrain the response of the beamformer such that the signal in the desired direction is enhanced while interfering signals and noise are blocked. The minimization of interferences and noise is achieved by choosing the beamformer weights such that the output power is minimized [32, 69, 76].

$$E\{|y|^2\} = h^T R h \tag{64}$$

where y is the beamformer output R is the correlation matrix and h is the filter. The minimization of (64) is expressed as:

$$\min_h h^T R h \qquad \text{subject to} \quad a(\theta)h = g^* \tag{65}$$

where a is the steering vector for the array and g^* is a complex constant. By solving (65) using the Lagrange multipliers the filter h can be obtained.

$$h = g^* \frac{a(\theta)R^{-1}}{a(\theta)h^T R^{-1} a(\theta)} \tag{66}$$

when $g = 1$, (66) is known as the Minimum Variance Distortionless Beamformer (MVDR), (also known as the capon beamformer). Here, the covariance matrix of the array output is used in the derivation of the LCMV and the MVDR Beamformer, but in reality the covariance matrix of the array output contains the target signals as well as the interfering signals and the noise, hence the minimization of the covariance matrix is an approximation. The covariance matrix that has to be used is the covariance matrix of interfering signals and the noise, which is not available in practice [77]. Due to this assumption the performance of the beamformer suffers. There has been several approaches proposed for the accurate estimate of the covariance matrix which include the Eigen space [78, 79] approach and the diagonal loading approach [80]. These approaches improve the accuracy of the covariance matrix estimation and hence improve the performance of the beamformer.

An alternative approach to the LCMV beamformer is the Generalized Sidelobe Canceller (GSC). The advantage offered by the GSC algorithm is that it offers a data independent solution to the LCMV beamformer and it provides a mechanism for changing a constrained minimization problem into an unconstrained form. The most well known GSC implementation, proposed in [81] is known as the Griffiths and Jim (GJ) beamformer. The basic idea proposed in [81] is to divide the filter of the LCMV method into two components operating on the orthogonal subspace, which are a conventional beamformer and a sidelobe cancelling part. The GJ beamformer is shown in Figure 13, the beamforming operation is divided into three main parts; a fixed beamformer; blocking matrix; and an adaptive filter. The fixed beamformer time aligns the array output and enhances the desired source signal. The fixed beamformer can be a filter and sum or a DS beamformer. The blocking matrix is a rejection filter that blocks the desired signal and passes interfering signals and noise. The adaptive filter processes the outputs from the blocking matrix based on the feedback from the output of the beamformer. The delayed signal from the fixed beamformer is then subtracted from the output of the adaptive filter. One of the drawbacks of this beamformer is leaking of the signal from the blocking matrix; several solutions have been proposed to limit the signal leaking [82]. A comparison of the performance of the different variations of the GSC beamforming algorithms can be found in [83], where the results show that the best in terms of perceptual quality is the transfer function GSC.

The other approach used in beamforming is the LMS approach proposed initially in [84] variations proposed by many authors. The basis of the LMS algorithm is to minimize the error between a desired signal $d(n)$ and filter output.

$$e(n) = d(n) - hx(n) \tag{67}$$

The LMS algorithm can be expressed as [84]:

$$h(n+1) = h(n) + \mu x(n)[e(n)]^{*} \tag{68}$$

where μ is the step size and, the goal of the beamformer is to minimise the Mean Square Error (MSE). The proposed method for minimizing the MSE in [84] is by using gradient based steepest decent method. By applying the steepest decent method the weight function can be expressed as [84]:

$$h(n+1) = h(n) + \mu[\nabla E\{e(n)^2\}] \tag{69}$$

where the gradient vector which is the partial derivative of the MSE function with respect to h and is expressed as [84]:

$$\nabla E\{e(n)^2\} = 2Rh - 2r \tag{70}$$

where R_{xx} is the covariance matrix and r_{xd} is the cross correlation matrix between $x(n)$ and the desired response $d(n)$. Here MSE is minimum when the gradient is equal to zero. The drawback of the MSE function is the desired signal is often an unknown.

A detailed study of the performance of speech enhancement of all beamformers discussed is presented in [28], where results show that variations of LMS and LCMV beamformers perform the best under large impulse responses while MVDR beamformers are robust at small lengths of impulse response for ULA's.

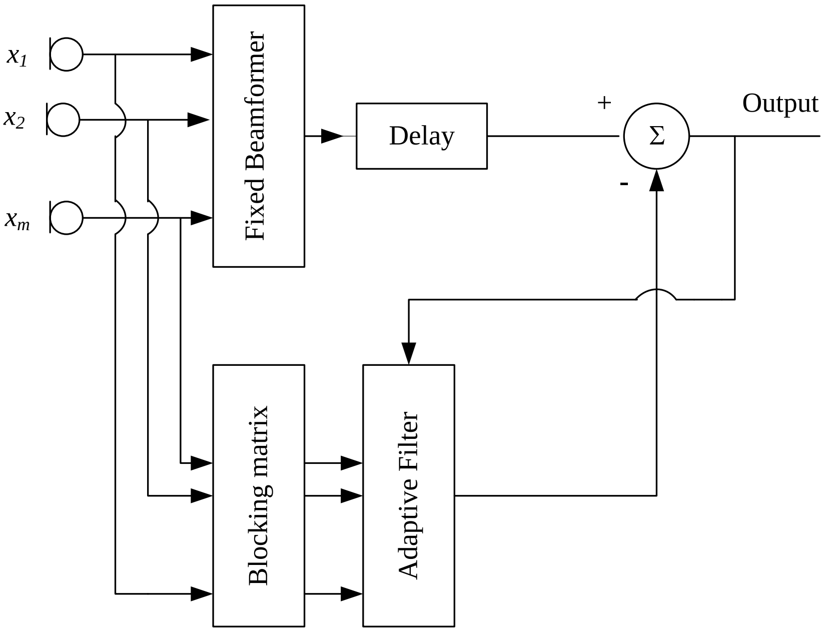


Figure 13: Block diagram of Griffiths and Jim Beamformer.

2.10 Speech Enhancement

The problem of enhancing noise corrupted speech is a well researched area. The methods proposed for speech enhancement include filtering, beamforming and source separation. Although these follow different approaches, in reality these three methods are related. When a single channel is considered, filtering is the best option and the other two methods does not produce significant results. However, studies have shown that the most effect way to enhance speech is based on multichannel recording like those from a microphone array [28]. Before looking at the different enhancement techniques, an introduction to speech signals and methods used in measuring the enhancement will be first discussed.

2.10.1 Human Speech

Speech signals are non-stationary that is the energy of the speech signals changes over time. However over time frames (10-30ms), the spectral characteristics of the speech can be considered as stationary. The process involved in the production of speech in human beings is extremely complex and involves the lungs, larynx and vocal tract (the organs in the mouth and the nasal cavity). The lung is the starting point of the speech production and the air in the lungs is exhaled though the larynx. In the larynx are the vocal cords (or folds), which oscillate to create sound. The closing and opening of the larynx is known as the glottal cycle or the pitch period and, the fundamental frequency of the speech is the reciprocal of the pitch period. The shape and muscular density of the larynx control the frequency of oscillation. The denser the muscle density in the larynx, the larger the pitch period and the lower the fundamental frequency, and this is why male voice is lower than female voice. The fundamental frequencies for males range from 60-150 Hz; whereas the fundamental frequencies for a female are 200-400Hz [85].

The sound vibrations created in the larynx passes through the vocal tract, which resonates to produce meaningful sounds. The shaping of the sounds from the resonation of the vocal tract is performed by the position of the tongue, teeth, jaws and lips (articulators). The frequency with which the vocal tract resonates is known as the formant frequency. The first four formants in human voice are the most important and are labelled as F0, F1, F2and F3.The F0 is known as the fundamental frequency and all the other formants are harmonics of the fundamental frequency. The F1 represent the sounds that require the mouth opening, F2 represent sound created by changing the position of the tongue and lips, and F3 is associated with front vs. back constriction in the oral cavity [85].

Human speech can be either voiced or unvoiced. Voiced speech occurs when the vocal folds are squeezed. An increase and decrease in the tension of the folds together with an increase and decrease in pressure causes the folds to open and close periodically, producing voiced speech. The sounds produced in this state are the vowels. The energy of vowels are higher than other sounds. Unvoiced speech is produced when the larynx is open, allowing the air to pass through with the wall of the larynx contracted to create a turbulent air flow known as aspiration. Unvoiced speech includes whispering sounds like “h”.

2.10.2 Spectral Representation of Speech

The frequency content of a speech signal can be represented by the spectral envelope of the speech spectrum as shown in Figure 14. The frequency contents that are most important for the speech signal are the first three formant frequencies. The speech energy in the spectrum is located below 1 kHz, and the peak is at 500Hz [85]. The different formant peaks in the spectrum play important roles in the identification of features of speech. The F0 formant is needed to identify different speakers while F1 and F2 formants are essential for the identification of the vowels and stop constants. When there is more than one speech source with the same F0 than the two sources will be indistinguishable. The difference in the F0 values is what distinguishes two speakers, hence the larger the difference in the F0s the easier it would be to distinguish the two speakers [86]. This concept is used in source separation algorithms which rely on the accurate identification of the F0 [87]. There are several methods that have been proposed for estimation of the F0 in single and multiple source scenarios [87].

One of the areas of speech enhancement is source separation when multiple speakers are overlapped. The difference in the formant frequencies between different speakers can be used in the separation of different speakers in mixed speech, this idea is used later in this thesis for sources separation.

2.10.3 The Human Auditory System

The Human auditory system can be divided into three main parts; the outer ear; the middle ear; and the inner ear. The outer ear consists of the pinna, the canal and the ear drum, while the middle ear consists of the three bones (ossicles) that are connected to the ear drum and the cochlear in the inner ear. The sound vibrations is channelled through the canal to the ear drum which vibrates, the vibrations of the ear drum are transmitted through the three bones in a lever action to the cochlear. The cochlear contains a fluid filled coiled cavity with two membranes known as the Resinner’s membrane and the Basilar membrane. The Basilar membrane varies in mass and stiffness at different regions and these regions have different resonant frequencies [87]. When the vibrations of the soundwave reach the cochlear the region in the cochlear that matches the frequency of the vibration resonates and these resonances are converted into neural activity through the hair cells and passed to the brain for processing.

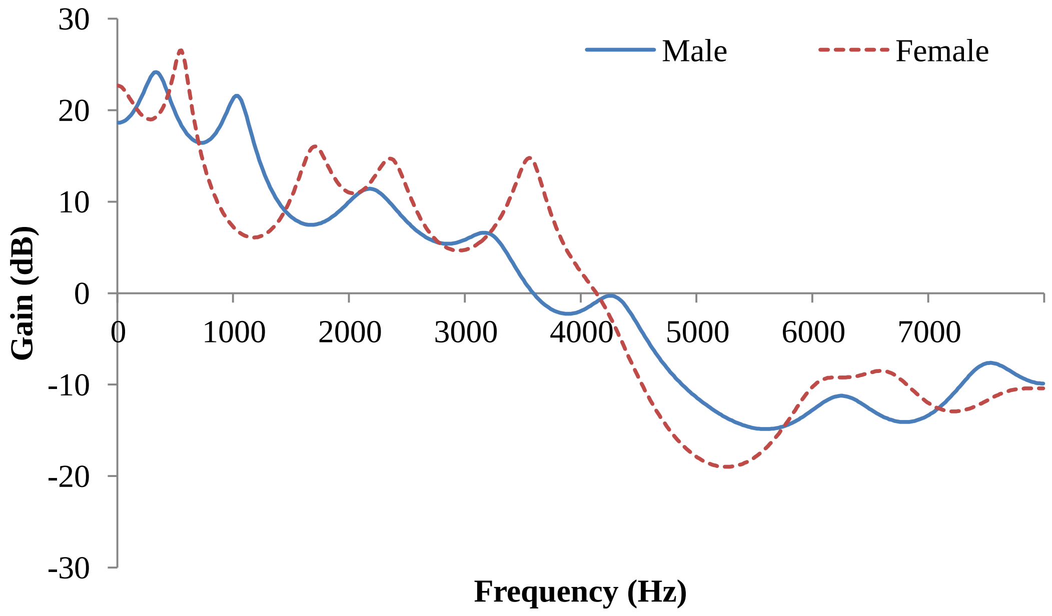


Figure 14: Envelope of spectra of vowel *"a"* for Male and Female speakers.

One of the most interesting aspects of human hearing is the ability to focus on a given sound in noisy areas. Due to the binaural structure of the ears, by moving the head and through selective filtering in the brain, human beings are able to filter out noise to some extent. Although the human auditory system has this ability, there are limits to which these abilities are true. When the competing sounds are too large then the desired sound is masked and cannot be distinguished. This is true when the frequencies of the competing source are close to each other.

The concept of masking can be explained as when one source has enough energy to hide the other, then the softer source is said to be masked by the stronger source. There are two types of masking which are:

- Simultaneous masking, which is a frequency domain phenomenon, occurs when weaker signal is made in audible by a stronger signal, which has a frequency close to that of the weaker signal.
- Temporal masking, which is a time domain phenomenon, occurs when a sudden high energy sound makes a low energy sound inaudible for a short period of time. Temporal masking can occur preceding the high energy signal or after the high energy signal. Since the effects of temporal masking last a short period, for enhancement of speech signals simultaneous masking is more applicable.

In general, tones are less effective in masking compared to broadband noise. When there is more than one speaker, especially when the speakers are of the same sex, then it is harder to separate the source from competing speakers. One of the mechanisms used by the human auditory system is to look for breaks in the mixed speech, but when the numbers of speakers are more than three, there are no breaks and sources are similar to stationary noise.

2.10.4 Types of Noise and Distortions

Noise can be considered as any sound that is not desired. Noise can be found in all environments and can be non-stationary and stationary. Stationary noise is any noise source whose energy remains constant over time. This includes noise from mechanical sources such as fans, air conditioning, moving vehicles, aeroplanes and coloured noise. The energy of nonstationary noise changes over time and examples of nonstationary noise are noise in parties and restaurants. Different noise types occupy different frequency ranges, those noise types that fall in the range of human speech (60-7000Hz) [85] are the most difficult to remove and the most destructive.

Distortion in speech can occur due to natural effects such as reverberation and echoes and manmade effects like filtering. The distortions that are introduced due to filtering, such as musical distortion, are caused by missing frequency components. Musical distortion is a common problem in subtractive enhancement algorithms [5, 88].

2.10.5 Speech Enhancement Algorithms

The aim of speech enhancement algorithms is to improve the perceptual quality and intelligibility of the speech which has been corrupted by noise, reverberation or distortion. The different classes of speech enhancement algorithms according to [85] are:

1. Spectral Subtractive algorithms – These algorithms are based on the idea that the noise in speech is additive, hence if the noise can be estimated it can be removed by simple subtraction of the noise from the noise corrupted signals.
2. Statistical based Algorithms – These include algorithms like the MSE, which are based on the statistics of the signals.

3. Subspace Algorithms – These algorithms are based on the concept of decomposing the signal space to a signal and noise, using methods such as Singular Value Decomposition (SVD).
4. Methods using source modelling.

At the start of this section, the three methods that were proposed for speech enhancement fit into the categories listed above. Most filtering algorithms that have been proposed for speech enhancement fall into the first and second category while beamforming fits into the second category and source separation algorithms are in the third category. A separate but important sub-area of speech enhancement is dereverberation. Most filtering using beamforming approaches perform some level of dereverberation, but dereverberation algorithms are generally regarded as a separate topic.

2.10.6 Filters for Speech Enhancement

The most commonly used filter for removing noise is the Weiner filter, which has been studied extensively and several variations of the original filter have been proposed. The Wiener filter is a subtractive algorithm which is not suitable for removing non-stationary noise. In Section 2.9.13, (64) gives the error between the desired signal and the output of the filter. The resultant filter from the minimization of (67) is the Wiener filter. Hence, a single channel implementation of the LMS based beamformer can be considered a Wiener filter. The Wiener filter will still have the same drawbacks as the LMS approach, since the noise source is unknown. In typical Wiener filter implementations, the noise is estimated from breaks in the speech or by using a section at the start of the recording to estimate the noise. In addition to the problem of estimating the noise, the Wiener filters suffer from the problem of musical distortions in the filtered speech due to removal of critical frequency components for speech.

If the Weiner filter is expanded to a multichannel case then it can be shown that the multichannel Weiner filter and the MVDR filter are identical [28]. The detailed derivation of the proof can be found in [28]. There are several variations of the Weiner filters proposed. One of the proposed variations is the Distortionless Wiener filter with psychoacoustic constraints as proposed by [85, 89]; this filter is of particular interest as it tries to address the problem of the distortions introduced by Wiener filters. In [85], a variation of the Wiener filter that is based on minimizing the distortions caused by noise is given. The performance of this filter in comparison to other variations of the Wiener filter is much better. A detail discussion of most types of Wiener filter can be found in [85].

2.10.7 Distortionless Wiener Filter

A noise corrupted speech signal can be expressed as:

$$\boldsymbol{y}(t) = \boldsymbol{x}(t) + \boldsymbol{n}(t) \tag{72}$$

where $\boldsymbol{y}(t), \boldsymbol{x}(t)$ and $\boldsymbol{n}(t)$ are the vectors of the noise corrupted speech, clean speech and noise respectively. If F is an N point DFT matrix, then (72) can be expressed in the the frequency domain as [85]:

$$Y(\omega) = F^H.\boldsymbol{y}(t) = F^H.\boldsymbol{x}(t) + F^H\boldsymbol{n}(t) = X(\omega) + N(\omega) \tag{73}$$

$$\hat{X}(\omega) = G.X(\omega) + G.N(\omega) \tag{74}$$

where $\hat{X}(\omega) = G.Y(\omega)$ is a linear estimation of the $X(\omega)$ and G is an $N \times N$ diagonal estimator. The error in the frequency domain can be derived according to [85] as :

$$e(\omega) = \hat{X}(\omega) - X(\omega)$$

Using (74), this reduces to

$$e(\omega) = (G - I).X(\omega) + G.N(\omega)$$

$$e(\omega) = e_x(\omega) + e_n(\omega) \tag{75}$$

where I is $N \times N$ identity matrix, and e_x and e_n are distortion terms. The energy of the distortions can be expressed according to [85] as:

$$e_x^2 = E\big(e_x^H(\omega).e_x(\omega)\big) = tr(E((G-I).X(\omega).((G-I).X(\omega))^H$$

$$= tr((G-I).F^H.R_{xx}.F.(G-I)^H) \tag{76}$$

where tr is the trace of matrix and similarly the energy in noise as:

$$e_n^2 = tr(G.F^H.R_{nn}.F.G^H) \tag{77}$$

where R_{xx} and R_{nn} are the autocorrelation matrices of the speech and noise. The distortion in speech can be minimized by solving the constrained optimization problem:

$$\min_G e_x^2 \tag{78}$$

$$\text{subject to:} \frac{1}{N} e_n^2 \leq \delta \tag{79}$$

where δ is a positive number corresponding to the minimizing threshold for noise. The result is the minimization of the distortion of speech in the frequency domain and maintaining the energy of the residual noise below the threshold δ [85]. In [85] the Lagrange method is used to solve the minimisation problem resulting in:

$$G.(F^H.R_{xx}.F + \mu.F^H.R_{nn}.F) = F^H.R_{xx}.F \tag{80}$$

By making the following assumptions (80) can be solved to give a gain function that minimizes the noise. G is a diagonal matrix, and $F^H.R_{xx}.F$ and $F^H.R_{nn}.F$ are asymptotically diagonal, and the autocorrelation matrices are Toeplitz [85]. The diagonal of $F^H.R_{xx}.F$ and $F^H.R_{nn}.F$ are the power spectrum components $P_{xx}(\omega)$ and $P_{dd}(\omega)$ of the clean and noise vectors [85]. The gain function for each frequency component can be expressed as [85]:

$$g(k) = \frac{P_{xx}(\omega_k)}{P_{xx}(\omega_k)+\mu P_{nn}(\omega_k)} \tag{81}$$

where μ is the Lagrangian multiplier which is found according to [85]:

$$\mu = \begin{cases} \mu_0 - \frac{SNR_{dB}}{s} & if \quad -5 < SNR_{dB} < 20 \\ 1 & if \quad SNR_{dB} \geq 20 \\ \mu_{max} & if \quad SNR_{dB} \leq -5 \end{cases} \tag{82}$$

where μ_{max} is the maximum allowable value of μ, $\mu_0 = \frac{(1+4\mu_{max})}{5}$ and $s = \frac{25}{(\mu_{max}-1)}$ and the value of SNR is calculated as [85]:

$$SNR = \frac{\sum_{k=0}^{N-1} P_{xx}(\omega)}{\sum_{k=0}^{N-1} P_{nn}(\omega)} \tag{83}$$

The enhanced frequency spectrum $\hat{X}(\omega_k)$ can be obtained from $\hat{X}(\omega_k) = g(\omega_k).Y(\omega_k)$. The distortionless approach described above can be further enhanced by incorporation of a perceptual filter [85].

The approach proposed in [89] is based on the perceptually weighted error criteria used in low rate speech coders, which takes advantage of the masking properties of the human auditory system [85]. As explained earlier, the human auditory system cannot distinguish between two sounds when one has higher energy that the other. This concept of masking is used in speech coders to mask the quantization noise near the high energy regions of the spectrum. By exploiting the masking characteristics, a filter can be designed which places a higher emphasis on the spectral valleys of the spectrum where the noise is audible [85, 89, 90]. The filter that is used is based on the analysis-by-synthesis filter used in Linear Prediction (LP) modelling of speech. The filter is expressed as:

$$H(z) = \frac{1-\sum_1^p a_k z^{-1}}{1-\sum_1^p a_k \gamma^k z^{-1}} \tag{84}$$

where p is the order of prediction and a_k are short term prediction coefficients and γ $(0 \leq \gamma \leq 1)$ is a parameter that controls the error in the formant regions. The plot of the spectra for the (84) is shown in Figure 15. From Figure 15 it can be seen that the filter places more emphasis on the spectral valleys than on the formant peaks. In [90] the constraints present in (79) are replaced by perceptually weighted noise, and this perceptual weighting will make the noise inaudible. The noise energy in (77) can be expressed in terms of perceptually weighted noise as [90]:

$$e_{pn}^2 = tr(W_f . E[e_n e_n^H] W_f^H) \tag{85}$$

where W_f is the perceptual weighting matrix based on the perceptual filter (84) and is given as [90]:

$$W_f = \begin{bmatrix} H(0) & 0 & \cdots & 0 \\ 0 & H(\omega_0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & H((N-1)\omega_0) \end{bmatrix} \tag{86}$$

where $\omega_0 = \frac{2\pi}{N}$. Similar to the derivation of the gain functions for the distortionless Wiener filter, the gain function that is based on perceptual weighting can be obtained. The perceptually based gain function is expressed as [90]:

$$g_p(k) = \frac{P_{xx}(\omega_k)}{P_{xx}(\omega_k)+\mu|H(\omega_k)|^2 P_{nn}(\omega_k)} \tag{87}$$

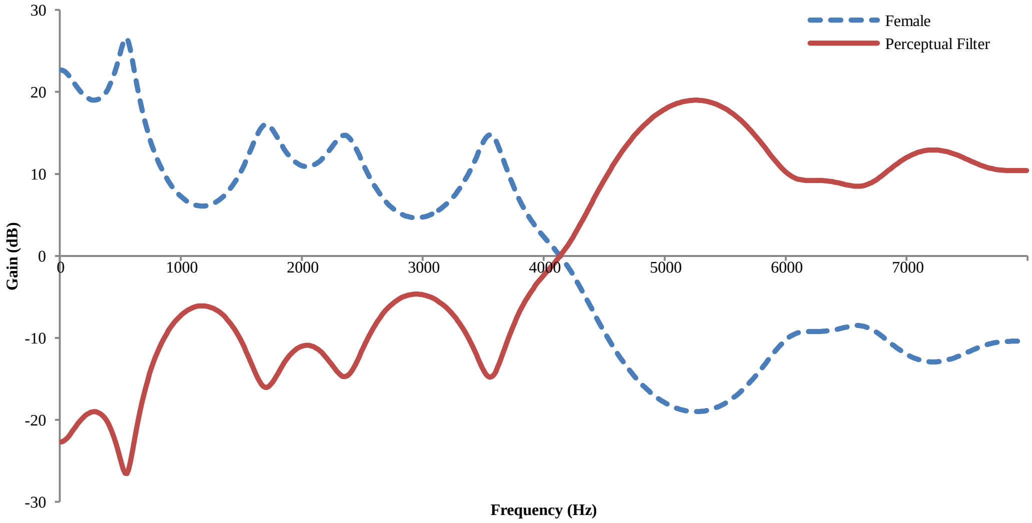


Figure 15: The plot of envelope of spectra for perceptual filter of (84), and the LP spectra for a female speaker.

The gain function from (87) can then be used to perceptually filter the noise corrupted speech signals. The only drawback of these algorithms is that since they rely on the knowledge of the noise covariance matrix; accurate estimates of noise may not be available. Hence, as with all other functions that rely on these unknowns, an estimate of the noise and speech has to be made. In addition to the noise estimates, the function relies on the accurate estimate of the LP spectra.

In addition to the Wiener filter, another popular filter used in speech enhancement is the Kalman filter. The Kalman filter is a Minimum MSE recursive estimator for a noise corrupted non-stationary signal. There are several variations of the Kalman filter for speech enhancement and most of these filters offer good quality in enhancing noise corrupted speech [91-93].

2.10.8 Speech Enhancement using Beamforming Techniques

A closely related topic to filtering is beamforming. In some aspects, beamforming offers more advantages compared to single channel filtering. These include the ability to steer the beam to any desired direction, and the use of multiple channels allows for uses of spatial and spectral information, whereas a single microphone system will only contain the spectral information. In addition to this, beamformers have been shown to offer some level of dereverberation in reverberant conditions. The constraints used in most beamformers are similar to those used in the filtering process. As discussed before it can be shown that the MVDR beamformer is identical to the multichannel Wiener filter [94]. Several authors have proposed different variations of beamformers for speech enhancement with good results [83, 95-98] .

2.10.9 Blind Source Separation Algorithms for Speech Enhancement

Blind Source Separation (BSS) has been one of the most difficult problems in speech signal processing, especially in reverberant conditions with multiple sources. The use of BSS for speech enhancement has also been proposed. The BSS algorithms can be broadly divided into time domain and the frequency domain, and further divided in algorithms for separating instantaneous mixtures and convolutive mixtures. The majority of early work done on this subject was based on instantaneous mixing models, which does not represent real world mixing models, as most environments are reverberant. The performance of these algorithms when applied to recordings from reverberant rooms suffer, especially when the levels of reverberation are high.

The instantaneous mixing model for *m* sources captured using *j* microphones can be represented as[99]:

$$x_j(k) = \sum_{i=1}^{m} h_{ji}\, s_i(k) + n_j(k) \tag{88}$$

where h_{ji} is the mixing model for the j^{th} sensor and i^{th} source and n_j is the noise at the j^{th} sensor. Here, only one instance of each source is added together. The convolutive mixing model can be expressed similar to (88) but since there is an infinite number of time delayed instances of each source due to multipath effects, the convolutive mixing model can be expressed as [99]:

$$x_j(k) = \sum_{l=-\infty}^{\infty} \sum_{i=1}^{m} h_{jil}\, s_i(k-l) + n_j(k) \tag{89}$$

where l represents the delayed versions of the source at each microphone. The multipath effect due to reverberation causes the mixed signal to be more complex than the non reverberant case and algorithms designed for reverberant conditions must be able to address both spatial un-mixing and the temporal changes that have been introduced into the mixing matrix.

There are some characteristics of speech signals that allow BSS algorithms to effectively separate mixed signals. These include [99]:

- Speech signals as described before are represented between the frequencies of 50 to 4 kHz

- The speech signals are non-stationary and amplitude modulations are largely responsible for this characteristic.
- It can be assumed that in a group different speakers will be located in different positions.
- Each speech signal has a unique temporal structure over short time frames.
- Speech signals are quasi- stationary for small time durations but non stationary over longer periods.

The successful BSS algorithms for speech separation use more than one of these features of speech, while it is possible to design a system that utilizes only one of these features. The most widely used source separation algorithm is the Independent Component Analysis (ICA) [100]. The original ICA algorithm was designed to separate instantaneous models; the convolutive fast ICA algorithm proposed in [101, 102] addresses the convolutive case. A detailed derivation of the ICA algorithm can be found in [103]. The basic assumptions that are made in the derivation of the ICA algorithms are those that have been listed above, and in particular the statistics of the different recordings are different. One of the important mechanisms relied upon by many BSS algorithms to get this statistical difference is the recordings are done using spatially distributed microphones.

There are many other methods that have been proposed for BSS and details of their implementation can be found in [99]. A comparative study between the different types of source separation algorithms found that ICA and its family of algorithms were the most efficient in terms of speed while the J. F. Cardoso’s ICA algorithm (JADE) algorithm showed the best performance in simulated cases in terms of SNR results [104].

2.10.10 Dereverberation of Speech Signals

The effects of reverberation can be considered as both required to some extent and a source of degradation when present at high levels. The effects of reverberation in moderate levels are required to make the sound more natural. This is seen when a person enters an anechoic chamber, where there is a disturbing feeling, as the human ear is designed to take advantage of the reflections for source localization and to control the loudness and pitch while speaking. This can be considered as the feedback mechanism that is needed by the human vocal and auditory system. The two most common perceptual effects related to the reverberation are the box effect and the distant taker effect. The box effect can be described as the sound coming from more than one direction at different times and adds the effect of spatialness and the distant talker effect is when the sound is seen to be coming from a distant point.

The destructive effects of the reverberation are when there are too many reflections and when the time taken for the reflection to die off is too high. It is in these cases that dereverberation is essential. There are several methods that can be used for dereverberation: they include beamforming methods; speech enhancement methods; and blind system identification and equalization methods, where the acoustic impulses are identified blindly and then used to design an equalization filter that compensates for the effect of acoustic impulse responses [29]. Most dereverberation algorithms are based on models that require the room impulse models which in practice is not available in most instances and it is difficult to obtain. The first two approaches described can provide some level of dereverberation, but exact dereverberation can be provided by the third approach [29]. The implementation of such algorithms are not practical due to high computational complexity and sensitivity to noise [29]. There are many algorithms that have been proposed for dereverberation of speech signals [29]. In particular those methods based on LP Spectra are of interest as these techniques are based on perceptual models, and so a better outcome can be expected from them.

2.10.11 Linear Predictive Coding (LPC) Based Dereverberation Approaches

The Linear Predictive Coding (LPC) based speech enhancement has been described before which has been used for removal of noise while the perceptual quality of the filtered speech is maintained. In [105, 106], it has been shown that the effects of the reverberation in speech are mainly on the prediction residual, especially in the case where recordings are made using microphone arrays.

An enhanced version of the LPC residual signal is used in synthesizing a speech signal with reduced reverberation from the output of a filter employing the LPC coefficients of the reverberant speech. One benefit of these algorithms is that no knowledge of the room impulse response is required for the dereverberation.

2.10.12 LPC of Speech

LPC of speech is used in speech coders to model the perceptually important spectral characteristics and quantise, transmit parameters of this model to facilitate efficient bit rates. A speech signal s(n) can be expressed in terms of a p^{th} order linear predictor as [29, 94]:

$$s(n) = \sum_{i=1}^{P} a_i s(n-i) + e(n) \tag{90}$$

where a_i are the prediction coefficients and $e(n)$ is the prediction error. The all pole LPC analysis filter from the LP coefficients of (90) is given as:

$$H(z) = \frac{1}{1+\sum_{i=1}^{P} a_i z^i} \tag{91}$$

The problem of obtaining the LP coefficients is solved by minimizing the MSE of the prediction error. The MSE function used is:

$$J = E\{(s(n) - \sum_{i=1}^{P} a_i s(n-i))^2\} \tag{92}$$

The error is minimized by setting the derivative J to zero with respect to each LPC coefficient:

$$\frac{\partial j}{\partial a_i} = 0 \tag{93}$$

The result of (93) is a set of p linear equations known as the normal equations and given as:

$$\begin{bmatrix} r_{ss,0} & r_{ss,1} & \cdots & r_{ss,p-1} \\ r_{ss,1} & r_{ss,0} & \cdots & r_{ss,p-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{ss,P-1} & r_{ss,p-2} & \cdots & r_{ss,0} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} r_{ss,1} \\ r_{ss,2} \\ \vdots \\ r_{ss,p} \end{bmatrix} \tag{94}$$

where $r_{ss,i} = E[s(n)s(n-i)]$ is the autocorrelation of the $s(n)$ for the i^{th} lag. The least square optimum estimates of the LP coefficients are given as:

$$\boldsymbol{a} = \boldsymbol{R}_{ss}^{-1}\boldsymbol{r}_{ss} \tag{95}$$

A common method used to solve (101) is the Levinson Durbin algorithm, (detailed derivation of the LP coefficients can be found in [94]). The derivation given above is for a single channel. There are several methods that can be used for obtaining the LP coefficients for multichannel case, which will be discussed next.

2.10.13 Multichannel LP Analysis

The LP coefficients of multichannel recording can are obtained using the following methods.

1. Beamforming of the multichannel recordings
2. Using the averaged autocorrelations of all the channels to calculate the LP coefficients
3. Using the Multivariate Auto Regression

The beamformer used in this approach is the DS beamformer. The aim of the beamforming operation is to combine the individual channels into single channels which can then be used to obtain the LP coefficients. The problem of obtaining LP coefficients from the DS beamformer is that the microphones are spatially distributed and there is a significant difference between the two channels.

2.10.14 The Averaged Autocorrelation of Channels

The averaged autocorrelation method can be expressed as:

$$\hat{R}_{ss} = \frac{1}{M}\sum_{m=1}^{M} R_{ss,m} \tag{96}$$

$$\hat{r}_{ss} = \frac{1}{M}\sum_{m=1}^{M} r_{ss,m} \tag{97}$$

where $\hat{R}_{ss}$ and $\hat{r}_{ss}$ are the averaged autocorrelation function, and M is the number of channels. According to [29] this method provides the best estimate for the LP coefficients compared to the single channels case and the DS beamformer.

2.10.15 The Multivariate Auto-regression Algorithm (MVAR)

The MVAR algorithm has been proposed by many for obtaining the multichannel LP coefficients for multichannel speech and audio coding. The multichannel signal can be expressed as:

$$\boldsymbol{x}(n) = [x_1(n) \quad x_2(n) \quad \ldots \quad x_M(n)] \tag{98}$$

The prediction error can be expressed as:

$$\boldsymbol{e}(n) = \boldsymbol{x}(n) - \sum_{i=1}^{P} \boldsymbol{A}_i\, \boldsymbol{x}(n-i) \tag{99}$$

$$\boldsymbol{A}_i = [A_1 \quad A_2 \quad \ldots \quad A_P\,]$$

where A_i is a $M \times M$ matrix. Here, a key difference to the single cannel case is that each LP coefficient is an $M \times M$ square matrix. The MSE is minimised using the multichannel Wiener Hopf equation and the LP coefficients are obtained using the Levinson-Wiggins-Robinson Algorithm. Since the LP coefficients are in blocks the total number of LP coefficients is $M \times PM$. Each A_i block coefficient matrix contains LP coefficients from autocorrelation and the LP coefficients from cross correlation of the multichannel signals. According to [107], compared to the complexity of the algorithm, very little coding gain is achieved. Furthermore, when the two channels are exactly the same or when one channel is zero, the problem of matrix singularities arises [107].

2.10.16 LPC Based Dereverberation Methods

There are many methods that employ the LPC residual filtering for dereverberation. The main goal is to apply different filters to the prediction residual such that de-reverberated speech can be obtained. One of these methods is the Spatiotemporal Averaging Method for Enhancement of Reverberant Speech (SMERSH) [108]. The SMERSH algorithm is made up of four major parts [29]:

- Time alignment of the signals to emphasise the direct path components.
- Detection of Glottal Closure Instances (GCI) such that the prediction residual can be segmented into individual larynx cycles.
- Averaging of the larynx cycles to obtain an enhanced larynx cycle.
- Voiced/unvoiced and silence detection

The SMERSH algorithm is based on using the information from the GCI to suppress the uncorrelated features of the prediction residue. The identification of the GCI’s is performed using the multichannel Dynamic Programming Phase Slope Algorithm (DYPSA) [108]. A detailed derivation of the multichannel DYPSA and the SMERSH algorithms can be found in [108].

Other LPC based dereverberation algorithms include the Regional Weighting Function and Weighting Function Based on Hilbert Envelopes [106] [109], and Wavelet Extreme Clustering [110] to name a few.

2.11 Measuring the Amount of Enhancement for Speech Signals

The literature on speech enhancement generally uses the measure of Signal to Noise Ratio (SNR) to measure the performance of the enhancement the techniques. SNR gives a good indication as to how well the level of noise has been attenuated. What it fails to indicate is how well a given system performs in a perceptual sense. The perceptual quality of the output of a system can be measured from either listening quality, speaking quality or conversational quality. In this work only listening quality will be examined. The psychological factors that are involved in determining speech quality include:

- Naturalness – Does the recoding sound natural?
- Intelligibility – Can the listener understand all the words correctly?
- Loudness – Is the recording at comfortable level?

The combination of all these factors determines the overall quality of speech. But these factors can be individually used to measure a specific aspect of a system. As an example, the intelligibility test can be used in source separation to measure how well a given source has been separated from the mixed recording, but if asked to measure all together the test subjects may rank quality based on the wrong source. Hence, it is important to choose the correct measure for the correct test.

Testing can be carried out at two levels, which is using a set of words which have no relation to each other and sounds similar and asking the listeners to identify the words (Diagnostic Rhyme Test), or sentence level tests. For the sentence level testing, the sentences are chosen such that they are phonetically balanced and not meaningful [111-113], meaning that the test subjects will not be able guess the words in the sentence.

The measurement of speech quality can be divided into either subjective or objective testing. Subjective testing involves the using of a group of listeners who are asked to rank the quality. The objective testing involves the use of a computer program to emulate a human listener and rank accordingly. The standard used for testing the transmission quality for communications systems is outlined in the International Telecommunications Unions (ITU) recommendations [114, 115].

2.11.1 Subjective Tests for Speech Quality

The test carried out using human subjects who are generally non expert listeners who are native speakers of the language of which the testing is carried out. One of the problems of a listening test is that after a while the listeners may get used to listening, hence they may score high, and also the listeners may get bored if there are too many files. These problems can be avoided by limiting the number of test sentences [116].

The listeners are played a test recording once and asked to rank based on intelligibility, loudness and naturalness. The way the recordings are played and the content of the recordings does have an impact on the ranking by a test subjects, hence in general the recordings are played in random order unless the test is carried out to compare the performance of different algorithms, in which case a set of files are played randomly ordered. Depending on the objective of the test, a range of values is set. For listening quality, the five level scoring system in Table 1 is used.

Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table 1: The MOS scale.

A Mean Opinion Score (MOS) for the algorithm is generated from the average of the scores from all the listeners. A closely related scale is used to measure the different distortions of the algorithms under test. The listeners are played two files and based on the first file the listeners rank the second file on the level of degradation. This is known as a Degradation Mean Opinion Score (DMOS). The scale used in the DMOS test is shown Table 2.

Inaudible	5
Audible but not annoying	4
Slightly annoying	3
Annoying	2
Very annoying	1

Table 2: The DMOS scale

A third form of subjective test is the Multi-Stimulus test with Hidden Reference and Anchor (MUSHRA) test as used to test audio quality. Here, the listeners are given a reference, and several other files. The listeners can listen to the files any number of times and can make comparisons with the reference. The listeners are then asked to give a score between 0 and 100 for each file except the reference file. There are several other testing systems, the details of which can be found in [94]. The most widely used testing system for speech listening quality is the MOS test, due to its simplicity.

2.11.2 Objective Tests for Speech Quality

The objective tests that can be used to evaluate speech quality are:

- The Signal to Noise Ratio (SNR)
- Signal to Interference Ratio (SIR)
- Signal to Distortion Ratio (SDR)
- Signal to Reverberation Ratio (SRR)
- Log Spectral Distortion (LSD)
- Itakura Saito Distance (ISD)
- Perceptual Evaluation of Speech Quality (PESQ)

The SNR, SIR, SDR and the SRR each measure a given distortion in speech. In contrast, the PESQ is based on the overall quality of the speech signals. The speech qualities in terms of individual measures are not suitable to completely describe the distortions, but a combination of these measures can give an accurate indication of the level and type of distortion.

Let the speech signal be represented as:

$$x(t) = s(t) + n(t) + i(t) + dis(t) \tag{100}$$

where $s(t)$ is the target signal, $n(t)$ is the noise in the channel, $i(t)$ represents interferences and $dis(t)$ is the distortion. The ratios listed from (1-3) can be defined as [117]:

$$SNR = 10log\frac{\|s(t)+i(t)\|^2}{\|n(t)\|^2} \tag{101}$$

$$SIR = 10log\frac{\|s(t)\|^2}{\|i(t)\|^2} \tag{102}$$

$$SDR = 10log\frac{\|s(t)\|^2}{\|n(t)+i(t)+dis(t)\|^2} \tag{103}$$

If the signal in (100) is redefined in terms of the direct path signal and the reverberant signal, then $x(t)$ can be expressed as:

$$x(t) = s(t) + s_d(t) \tag{105}$$

where $s_d(t)$ is a delayed version of the source signal $s(t)$. The SRR can be defined as [118]:

$$SRR = 10log\frac{\|s(t)\|^2}{\|s_d(t)\|^2} \tag{106}$$

2.11.3Log Spectral Distortion and Itakura Saito Distance

Distortion in speech signals can be measured based on a comparison of the spectral envelope of the signals. These methods are generally used in speech coding applications. If $f(\omega)$ is the spectral density of the speech signal, then the LSD can be defined as [119]:

$$LSD = \left\| log \frac{f(\omega)}{\hat{f}(\omega)} \right\| \tag{107}$$

where $\hat{f}(\omega)$ is the spectral density corresponding to the processed signal, here large values of LSD resemble higher distortion and smaller values resemble smaller distortions. A similar measure of distortion in spectral envelope is the ISD measure. The most widely used measure for the distortions between spectral envelopes between the two speech signals is the ISD [94]. It has been shown that the ISD can be used as an indicator for the subjective quality of speech. In [120], an enhanced version of the Itakura distance is presented, and it has been reported that if the ISD is less than 0.5 the difference MOS score is less than 1.6. The ISD between two signals can be expressed as [94, 119]:

$$ISD = \left\| \frac{f(\omega)}{\hat{f}(\omega)} - \ log \frac{f(\omega)}{\hat{f}(\omega)} - 1 \right\| \tag{108}$$

2.11.4 Perceptual Evaluation of Speech Quality (PESQ)

The Perceptual Evalution of Speech Quality (PESQ) is based on how the human ear detects the signals. Hence, the algorithms used in PESQ try to model the human ear as closely as possible. The PESQ models the human ear using filters that are a representations of the basilar membrane of the ear [121] using a three dimensional pattern representation in time, frequency and modulation frequency. The differences between the reference signal and test signals are performed using psychoacoustic models and translated into the MOS scale as an output [121-123]. The PESQ does not given an indication of the level of distortion caused by loudness loss, echoes and delays [94]. Furthermore, PESQ is designed to evaluate signals up to 8kHz bandwidth [115]. Hence, PESQ alone is not a good measure for speech quality; other measures have to be used together with PESQ to give an accurate estimate of the speech quality.

2.12 Conclusions and Summary

The work presented in this thesis is based on speech enhancement, DOA estimation and source separation using an in air acoustic vector sensor. The literature review presented in this chapter covers the basis that is needed to understand the work that is presented in this thesis. In this chapter, the basic principles of soundwaves, soundfields and sensors that are used for capturing soundwaves has been presented. These basic principles are critical in understanding how the AVS works and how the design of the AVS can be improved and how it captures sound sources.

The AVS is a co-located microphone array, hence it is important to understand the types of microphone arrays, how they are related and the design features that affect the performance of the microphone arrays are presented in this chapter. In addition to the design of microphones, the methods used in the processing of the outputs of the microphone arrays in general and how these approaches can be applied to AVS are discussed in this chapter.

In this chapter, the different methods used in enhancement of reverberant and noise corrupted speech for single and multichannel scenario is presented with emphasis on those algorithms that can be used with a co-located microphone array such as the AVS is presented. Finally objective and subjective methods that can be used to evaluate the performance of the enhancement algorithms is presented.

The work present in this chapter is used in the next chapter to examine the design of existing in-air AVS and to make improvements to the design of the AVS array in terms of directional and frequency response.

Chapter 3 AVS Design and Calibration

3.1 Introduction

The design of the AVS array is critical for the performance of the array in terms of accuracy and quality of the signals that is captured. The design goal here is to build an array capable of capturing high quality audio signals, with accurate directional information that can be used in the processing of the array outputs. Unlike the previous applications of an AVS, which were mainly for DOA estimation, here, the AVS is used for capturing and processing of speech signals which eventually will be transmitted through a communication channel.

An AVS has traditionally been used for DOA estimation of sources underwater and in air. The design of the AVS for use in underwater applications requires array sizes that have larger apertures than in air; this is due to the fact that the speed of sound is higher and the wavelengths are larger in water then for the same source in air [9]. The design of a large aperture array, especially for longer wavelengths is less complicated than designing arrays that are compact and designed for smaller wavelengths. The complication arises in designing the sensors that are capable of capturing the particle velocity accurately. The sensors that have been proposed for capturing particle velocity are hotwire anemometer, and the pressure gradient microphones. For applications in water where wavelengths are larger as those in sonar and seismic activity detection the size of these sensors are larger and easier to assemble, but for application in-air especially for speech and audio applications, the sensor sizes have to be small.

There are two different AVS designs that have been proposed for use in-air, one of which is the PU probe by Microflown [23], based on hotwire anemometers and the other is the native B format microphone array based on pressure gradient sensors. The difference between the two designs is price, size and the quality of recorded signals. The PU probe is extremely small and extremely expensive (approximately € 20,000 per probe) compared to the native B format microphones. In contrast to the PU probe, the native B format microphone array is capable of capturing soundfield at a higher quality which can be used for enhancement and reproduction. The price and the use of hotwire anemometers of the PU probe prohibit its use in mobile devices such as mobile phones and mobile computers. Hence, a more affordable, safe and attractive deign for the AVS is the native B format design.

In this chapter an AVS designed based on a native B format microphone array will be analysed [9] for the accuracy of DOA estimates based on the microphone response polar plots and DOA estimates. The frequency and directional response of the array will be used to measure its performance.

The methods for DOA estimation from an AVS array is presented in [7] and the effect of sensor placement on the accuracy of the DOA estimation has been presented in [6]. The relation between the sensor placement and the accuracy of DOA estimation can be used as a good indicator for designing AVS arrays to improve their performance.

Unlike DOA estimates, which depend on the design of the array and the performance of the algorithms used for DOA estimation, frequency and directional response of the microphones is only dependent on the design of the array and, hence is a better performance indicator than DOA estimates. Here, both methods will be used to evaluate the performance of different designs of AVSs.

The rest of this chapter is organised as follows: Section 3.2 presents the design of AVS based on the native B format array, which is an analysis of different types of AVS arrays and a study of the frequency response and directional response of the microphone capsules individually and attached to the arrays. Changes to the design based on the study will be presented and an evaluation of the performance of the new design based on microphone responses is presented. The output channels from the AVS array is presented in Section 3.3 followed by DOA estimation for measuring the performance of the AVS array, presented in Section 3.4. The outcomes are summarised in the conclusions of Section 3.6.

3.2 Design of AVS for In-Air Speech Signals

The criteria for designing the AVS for capturing speech signals in air can be summarised according to three features, which are: high quality recordings; accurate directional information; and an affordable price. As discussed in Section 3.1, a native B format array fulfils two of these criteria, that is, it is affordable and it has the potential to produce high quality recording. The rest of this section will look at existing native B

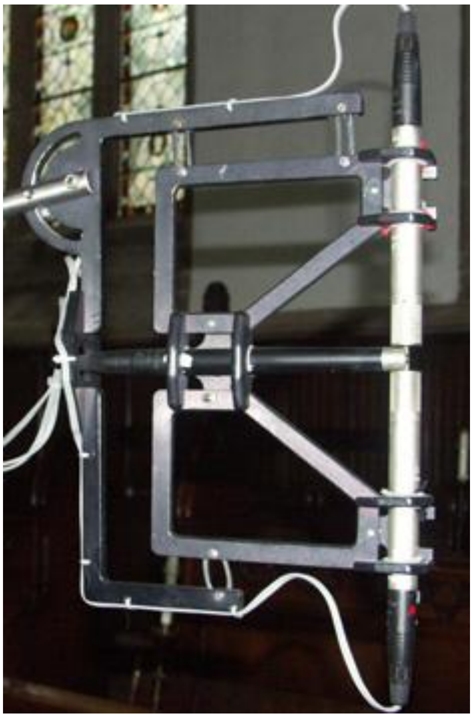
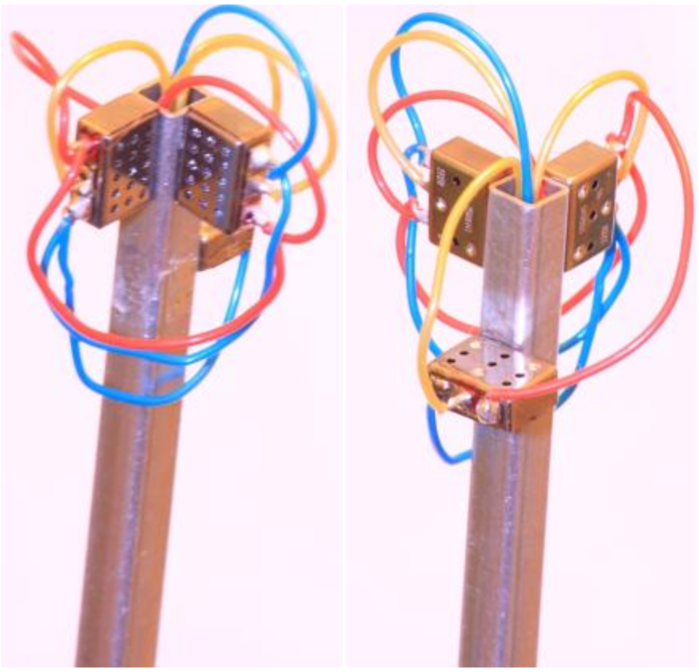


Figure 16: The Niumbus-Halliday Native B format array [124].

format microphone arrays and then evaluate the performance of the array based on frequency, directional responses and DOA estimates from the AVS arrays.

3.2.1 Different AVS Arrays

A native B format microphone, also know as a Nimbus-Halliday setup, was first proposed by Dr Jonathan Halliday (nimbus records) [125]. This microphone array consists of three microphones, which are two pressure gradient (figure-of-eight microphones) Schoeps bidirectional and a B&K omni-directional microphone as shown in the Figure 16 [125]. The array shown in Figure 16 uses commercially available microphones kept in place by a structure. The main drawback of this arrangement is the size of the physical array and the size of the microphones. Since one key requirement for accurate estimation of DOA from an AVS is co-location of the microphones, the Nimbus-Halliday arrangement does not fulfil this requirement. A more compact version of a native B format microphone is the Soundfield microphone, which has been discussed in detail in Section 2.8.7. The difference between these two arrays is the capsules used in the construction, the arrangement of the capsules, and way in which the directional signals are captured and formed. In comparison to the Soundfield microphone, the Nimbus-Halliday array is a better design as it does not need processing



(a) (b)

Figure 17 : The Lockwood Array (a) front showing the x and y sensors (b) back showing the Omni-directional sensor.

of the captured signals to form the x , y and w components, but due to its large size and the mechanism used to hold the array together, it is not a practical design for everyday use.

An array that is much more compact and truly co-located is presented in [9]. The array shown in Figure 17 is a two dimensional version of the array in [9] with x, y and o sensors, (the actual array presented in [9] is a three dimensional array with a pressure gradient capsule in the z direction). This array is comprised of four microphone capsules which are three Knowles NR-3158 pressure gradient sensors [126] and a Knowles EK-3132 omni-directional microphone [127]. Compared to the Nimbus-Halliday array, the aperture of microphone capsules used in the array of Figure 17 are extremely small and the structure holding the microphones in place is also extremely small.

The AVS array in [9] was tested with multiple sources, different beamforming algorithms for enhancement and with different numbers of sensors. The results presented in terms of SNR showed that the best performance is achieved when all the sensors on the array are used in the processing. The results showed that when there is only one interferer there is an improvement of 6 dB over unprocessed signals and an average of 4 dB improvements when there are 2 to 4 interferers. The results showed

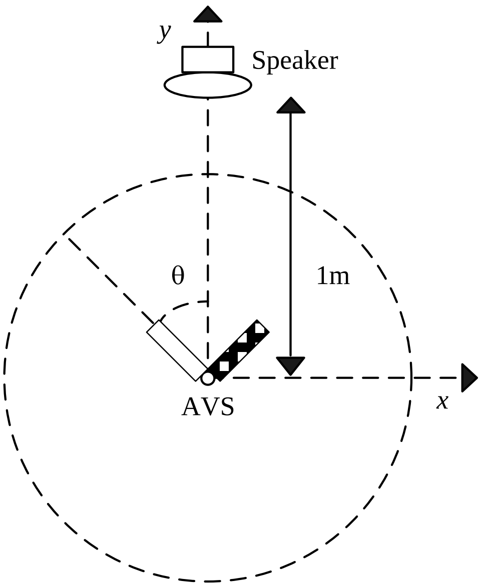


Figure 18: Setup for characterization of the AVS.

that the best beamformer for use with this AVS array is an enhanced version of the frequency domain implementation of the MVDR beamformer which is presented in [128]. The results presented in [9] are important in terms of beamforming, the work does not describe the response of the microphones used in the construction and results from DOA estimates are not present. The next section in this chapter will look at the microphones used in the construction of the AVS array of [9], which will be named as the Lockwood array for ease of discussion.

3.2.2 Response of the Capsules used in the Lockwood Array

The Lockwood array shown in Figure 17 has two types of microphone capsules used as described in previous section. In this section, the frequency response of these microphone capsules and the directional microphone response of the capsules will be presented. This information is important for identifying the behaviour of the microphone capsules once they are attached to the structure holding the AVS array together.

The study of the frequency and directional response is conducted in an anechoic chamber. The anechoic chamber allows monitoring the behaviour of the microphone capsules due to a single source without any reflections and echoes. The experimental methodology is described in the next section.

3.2.3 Experiments and Results

The experiments described in this section will enable the study of the response of the microphones that are used in the construction of the AVS array. There are two different tests that will be carried out:

1) The frequency response of the microphones and

2) The directional response of the microphones

The experimental setup of Figure 18 was used, where a single microphone is held in position by the connecting wires as shown in Figure 19, which were supported and passed through an aluminium square pole. The pole was mounted on a custom built rotating platform (to allow positioning of the microphones relative to the source) and a self powered speaker (Genelec 8020A) was placed in front of the microphone at a distance of 1 m with an elevation of 0 Degrees.

For the frequency response experiment an Exponential Sine Sweep (ESS) was played. For measuring the microphone directional response, a series of monotone signals each 2 seconds long and of equal energy were played with frequencies ranging from 100 Hz to 10 kHz. Recordings of 2 seconds long were made at 5 degree intervals and signals were sampled at 48 kHz.

3.2.4 Frequency Response of the Microphone

The frequency response of any microphone describes the behaviour of the output of the microphone to different frequencies. To obtain the frequency response of the microphone an impulse response measurement based on ESS is performed. The ESS can be expressed in the continuous time as [129]:

$$s(t) = sin\left[\frac{\omega_1 T}{ln\frac{\omega_1}{\omega_2}}\left(e^{\frac{t}{T}ln\left(\frac{\omega_2}{\omega_1}\right)} - 1\right)\right] \tag{109}$$

where T is time duration of the sweep, and ω_1 and ω_2 are angular frequencies corresponding to the start and stop frequencies. Here, the starting frequency is 1 Hz and stop frequency is 30 kHz. The sine sweep $s(t)$ is played from a loudspeaker 1m from the microphone array. To get the impulse response, the recorded signal is then convolved with the impulse response $f(t)$, which is the time-reversal of the test signal $s(t)$ and the impulse response is defined as [129]:

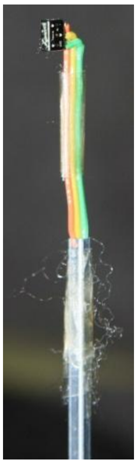


Figure 19: Single microphone used to get the frequency and directional response.

$$h(t) = r(t) \otimes f(t) \tag{110}$$

where $h(t)$ is the impulse response and $r(t)$ is the recorded signal and $\otimes$ is convolution. In total, two sets of recordings were made, one for the omni-directional sensor and one set for the pressure gradient sensor. Since the pressure gradient microphone has a directional response, impulse responses were measured for, 0, 45 and 90 degrees.

The plot for the frequency response of the omni-directional sensor is shown in Figure 20, where it is seen that the true response of the microphone is flat over the range from 50 Hz to 22 kHz. This frequency response is what is expected from an omni-directional sensor and the frequency response of the omni direction microphone remains constant for all source directions.

The frequency response plot of the gradient sensor is shown in Figure 21. Here, there are two important features of the microphone that has to be analysed, which are:

1. The frequency response plot shows that there is a boost in gain from 2 kHz,
2. The frequency response of the microphone maintains a similar pattern in gain levels for all azimuth angles tested but the gain levels change as the microphone is rotated in azimuth.

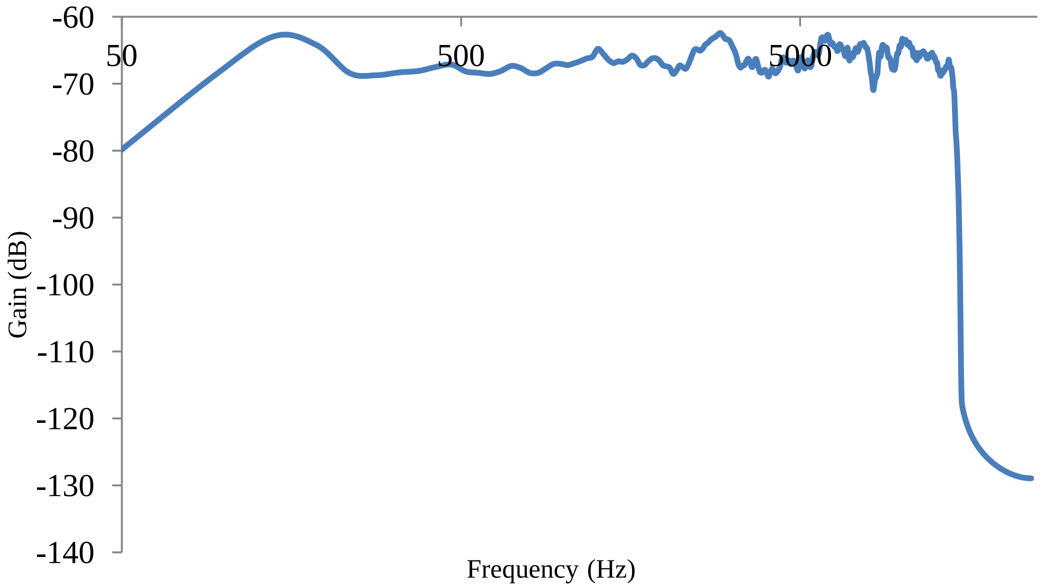


Figure 20 : Frequency Response of a Knowles EK 3132 Omni-directional microphone.

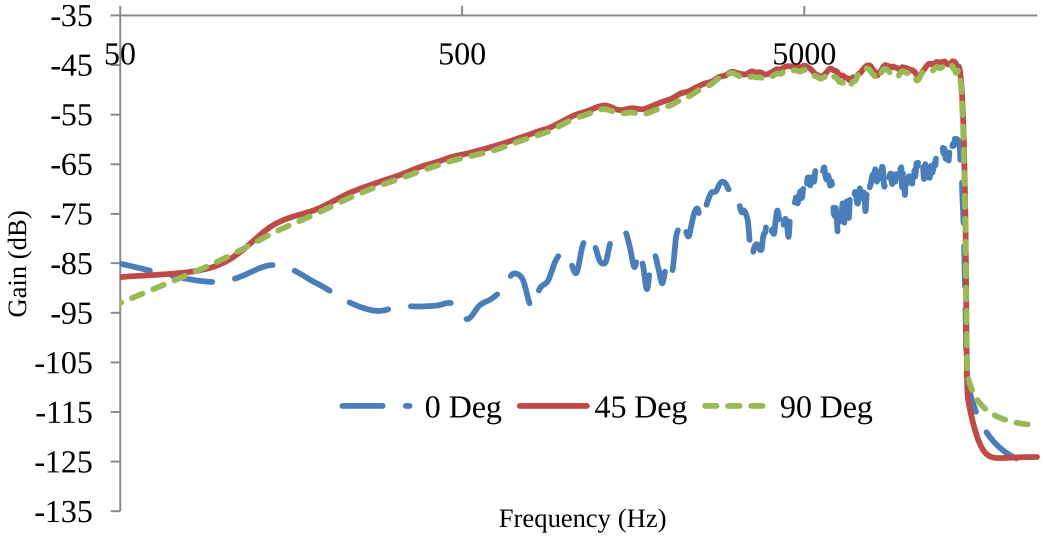


Figure 21: Frequency Response of a Knowles NR 3158 pressure gradient microphone.

The response of the microphone is seen to rise at a 6 dB/octave and falls at 22 kHz. This frequency is the frequency at which the wavelength of the soundwave is approximately equal to the separation between the front and back of the microphone which is 2.21mm. Hence, this is the maximum frequency to which the microphone is responsive and this agrees with the theory where the first null occurs when the wavelength of the soundwave is equal to the path around the microphone. The high frequency boost from 2 kHz is due to the effects of diffraction of the soundwave [26], which is a normal phenomenon in pressure gradient microphones. This high frequency boost can be compensated by a de-emphasise filter, which will be discussed later in detail. These results agree with the data sheets for the microphone capsules.

3.2.5 The Directional Response of the Microphones

The directional responses of the microphones are measured for only the pressure gradient microphone. The area around a pressure gradient sensor can be divided into four equal parts, which can be labelled as quadrant 1 to quadrant 4. The features of the directional response which are of interest are the symmetry of the plots in the four quadrants and the smoothness of the polar plots. The polar response was measured by finding the signal energy for each source location, which is expressed as:

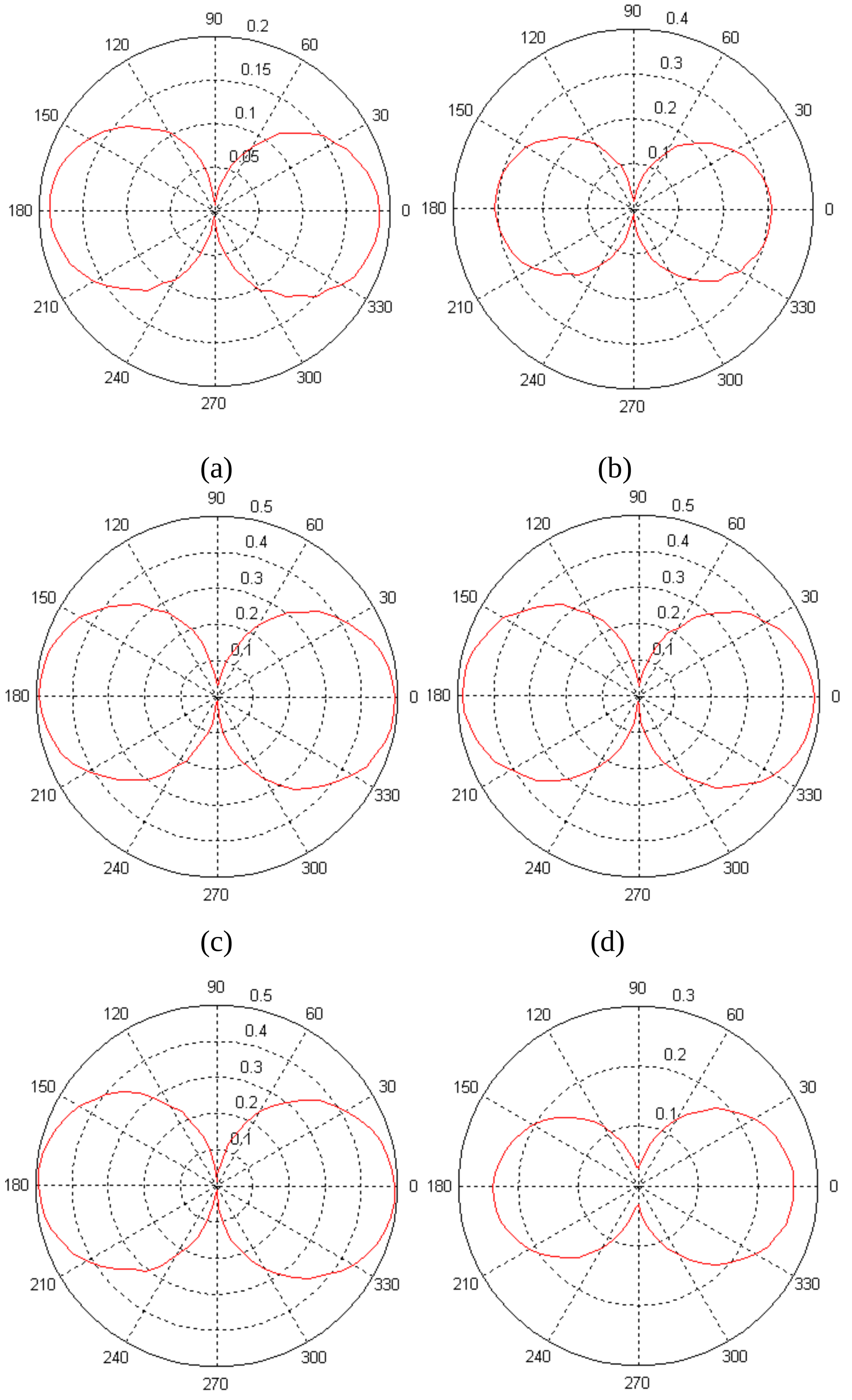
$$e = \sum_{n=1}^{N} |x_n|^2 \tag{111}$$

where e is the signal energy and N is the number of samples in the frame.

Figure 22 shows the polar plots of the directional responses of a pressure gradient microphone for selected frequencies from 100 Hz to 10 kHz. From the plots it can be seen that as the frequency increases, the maximum gain increases indicated from the increase in energy till 3 kHz, after which the gain is approximately constant for all frequencies up to 10 kHz. This is exactly as expected since the frequency response curves for the microphone shows that the gain is constant after 3 kHz.

The symmetry of all the plots is approximately the same and it can be seen that the polar plots are smooth. The symmetry indicates that there is no difference in the microphone pickup between the front and back. The smooth curves indicate that although there are some variations in the responses, overall there are no significant errors in the response due to imperfections in the construction of the microphone. Here, it is shown that without any additional support or other microphone capsules nearby the gradient sensors have a directional polar response which is approximately ideal.

In the next section, the frequency and the directional response of the microphones when they are mounted on the support structure of the Lockwood array will be investigated.



(a) (b)

(c) (d)

(e) (f)

Figure 22: The Directional Response of a Knowles NR 3158 pressure gradient microphone for different frequencies (a) 100 Hz (b) 500 Hz (c) 1 kHz (d) 3 kHz (e) 5 kHz (f) 7 kHz.

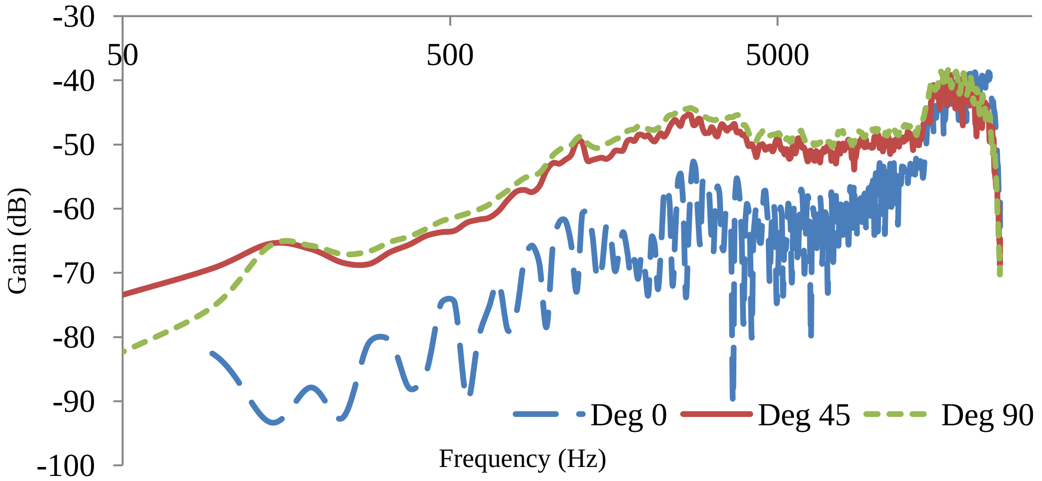


Figure 23: Frequency Response of the pressure gradient microphone in the Lockwood array.

3.2.6 The Frequency Response of the Sensors Attached to the Support

The frequency response of the microphone attached to support of the Lockwood array is investigated with the setup of Figure 18. The plot of the frequency response for 0 45, and 90 degrees is shown in Figure 23, where it can be seen that there is not much effect on the frequency response due to the support and the adjacent sensor. Here, the frequency responses of the x and y sensors on the array are averaged. The only significant change occurs in the value of the gain as the microphone is rotated in azimuth. In addition to the change in gain, the frequency response of the microphone is not smooth especially at low frequencies when the microphone is at 0 degrees to the source.

3.2.7 The Directional Response of the Sensors Attached to the Support

The directional response of the pressure gradient microphone without any support or interferers has been shown in Figure 22. Here the directional response of the pressure gradient microphone attached to an aluminium pole will be shown. The aluminium pole to which the microphones are attached is a square pole with a side of 2.5 mm, the thickness of the microphone itself is only 2.1mm and there are two microphones attached as shown is Figure 24. This arrangement makes the array extremely compact, where the approximate volume occupied by the microphones being 1 cm^3. The advantages gained by attaching the microphones to the square pole are:

1. The microphones can be positioned straight to the edge of the aluminium pole,
2. Two microphones can be place orthogonal to each other without errors in the angles between the microphones (the angle between the microphones have to be exactly 90 degrees),
3. The wires from the sensors can be managed such that they do not obstruct the microphones.

The effect of sensor placement on the performance of the DOA estimation was shown in [6] where it is shown that the asymptotic angular error depends on the array geometry.

The plots of the x and the y components of the directional response for selected frequencies between 100 Hz to 10 kHz are shown in Figure 25, from which it can be clearly seen that the symmetry of front, back, left and right lobes of the microphone are lost and smoothness of the plot is lost as well. As the frequency increases, the deformation of the directional response becomes more evident, especially frequencies above 6 kHz. It is seen that as frequency increases, the mid-part of the figure-of-eight grows wider. Furthermore, there is a distinct difference in size of the front and the back lobes of the figure-of-eight plots for the x and the y components. When compared to the single capsule case, the effect of the adjacent microphone is significant. These effects can be explained from (18), which is repeated here;

$$F(t) = \left[-jv_{par}\rho_0\omega d_{sep}\cos\theta\right]S \tag{18}$$

where v_{par} is the particle velocity, ρ_0 is density of air, ω is the angular frequency and d_{sep} and S is the front to back separation. It is shown that the driving force on the diaphragm of the pressure gradient sensor is a function of the particle velocity,

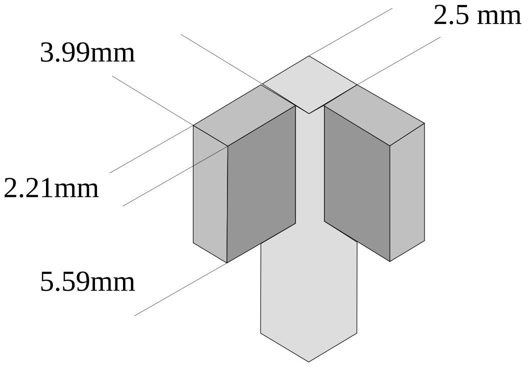


Figure 24 : The arrangement of microphones on a Lockwood array.

separation between the front and back of the diaphragm and the surface area. The output of the microphone can be expressed in terms of the driving force on the diaphragm as:

$$V_{out}(t) = aF(t) \tag{112}$$

where $V_{out}(t)$ is the voltage output from the microphone and a is the amplification from the internal circuitry of the microphone. Hence, any factors that affect the driving force on the diaphragm directly effects the microphone output.

The other factors such as density ρ_0, and the angular frequency ω remain constant. The neighbouring microphone capsule alters two very important variables in (18) which are the front to back separation and the surface area of the microphone. Both these variables are increased due to the metal support, and adjacent microphone, and the effect of the adjacent microphone is more significant than that contributed by the metal support.

The amount of distortions caused by the increase in the front to back separation and the surface area is a function of the DOA. As the source moves from quadrant 1 to quadrant 4 the values of the separation and the surface area change due to the change in shape as seen by the wavefront. The quadrant where the separation and the surface area is highest produces the larger lobes and quadrants where the separation and the surface area are small produces the smaller lobes.

In addition to these effects, the effect of shadowing by the adjacent microphone also contributes to the errors in the directional polar response of the microphone. The areas of the shadowing occur in the regions which are concealed to the soundwave the moment it comes in contact with the array; these regions are shown in Figure 31 for the Lockwood array.

Furthermore, there are the effects of reflection and diffraction due to the adjacent microphone and the metal support which also contribute to the error seen in the directional polar plots of the array, but the errors due to reflection and diffraction only start effecting at higher frequencies where the array size is close to ¼ of the wavelength. At higher frequencies, the waves bend and reflect more that at low frequencies; as a result the waves near 0 and 180 degrees for the *x* component and 90 and 270 degrees for the *y* component cause the errors in pressure difference. These errors are seen in figure-of-eight plot of Figure 25 (f).

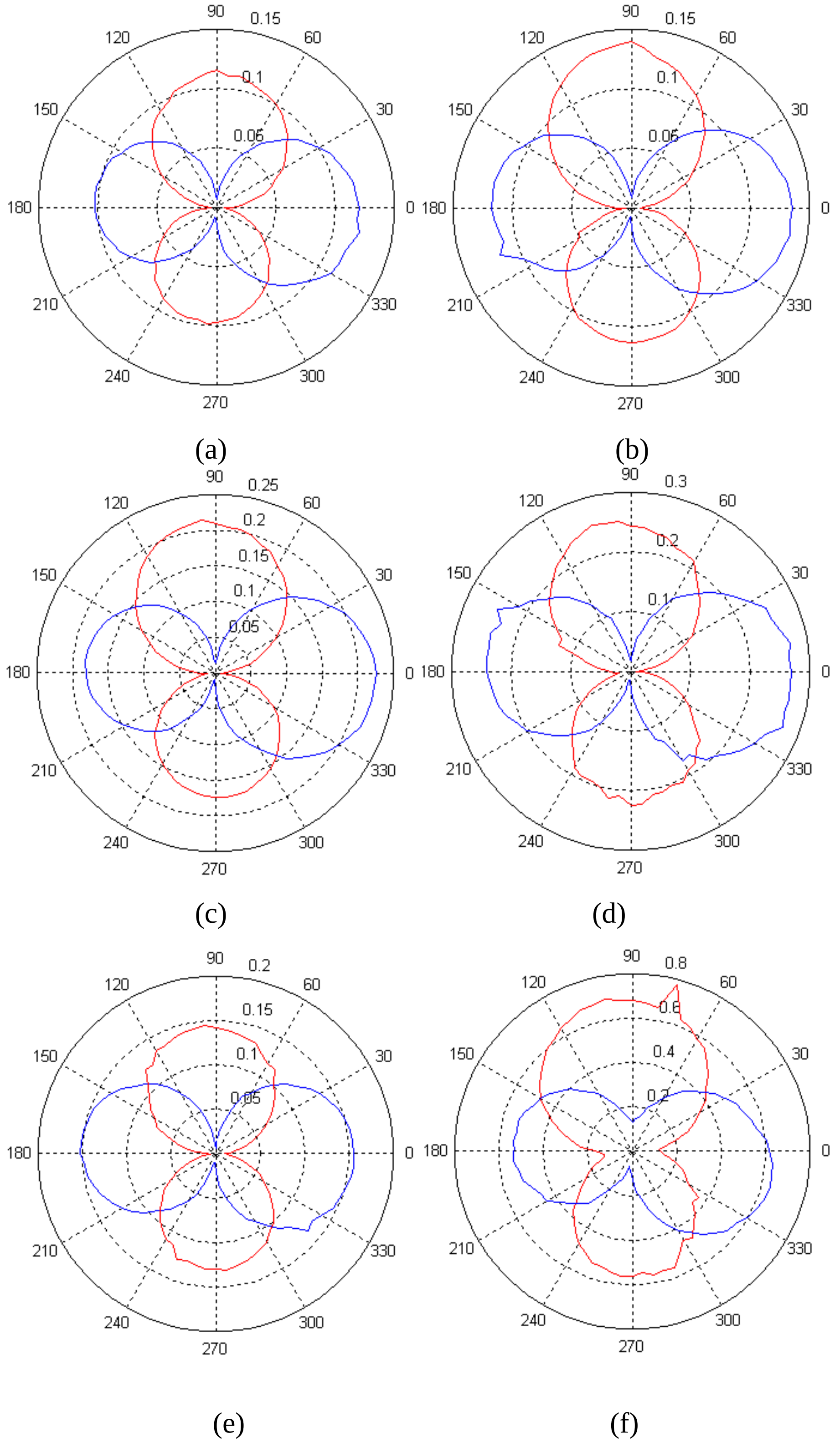
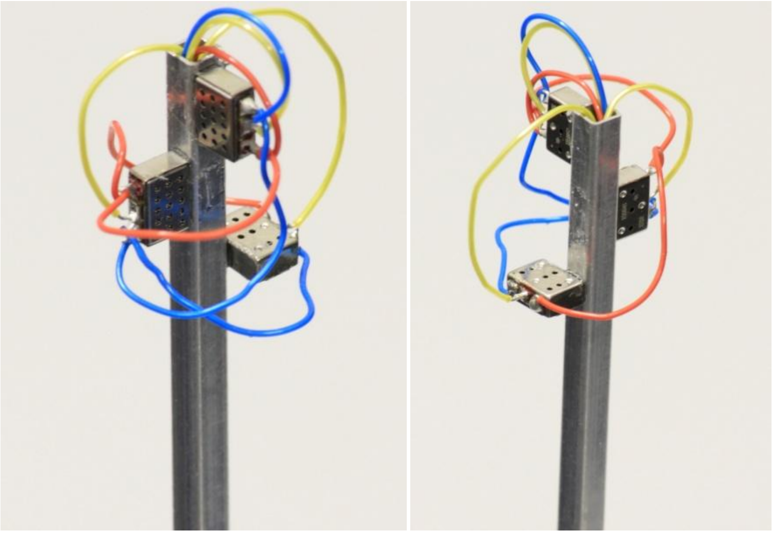


Figure 25: The Directional Response of pressure gradient microphones on Lockwood array. The x sensor is plotted in red and y sensor in blue (a) 100 Hz (b) 500 Hz (c) 1 kHz (d) 3 kHz (e) 5 kHz (f) 7 kHz.



(a) (b)

Figure 26 : The AVS II (a) front showing the *x* and *y* sensors (b) back showing the Omni-directional sensor.

Since the effects of the adjacent microphone have contributed to errors, a design change that improves the errors for the AVS is proposed in the next section.

3.2.8 The Offsetting of the *x* and *y* Microphone Capsules of the Lockwood Array

The errors in the directional response of Lockwood array presented in the previous section are due to the placement of the sensors adjacent to each other as described before. By offsetting the sensors such that the separation between the sensors are more than ¼ of the wavelength of the highest frequency, better results are expected. For descriptive purposes this array will be called AVS II. The proposed design change for the Lockwood array is shown in Figure 26, where the offset between the *x* and the *y* capsules is 0.5 cm which is approximately the ¼ wavelength of a 15 kHz wave. A study of the frequency and directional of the array responses was conducted similar to previous section.

3.2.9 The Frequency Response of AVS II

The frequency response of AVS II is performed for 0, 45 and 90 degrees in azimuth, with the ESS. The results of the frequency response of the microphones in AVS II are shown Figure 27. The frequency response of the microphones is seen to have less variation in the gain at 0 degrees compared to the results for the Lockwood array of Figure 23; furthermore the overall frequency response is smoother than the Lockwood array and is closer to the individual microphone response.

3.2.10 The Directional Response of AVS II

The directional response for selected frequencies similar to section 3.2.5 of AVS II is shown in Figure 29. The results show that there is an improvement in the symmetry of the polar plots. The left and right sides of the plots are more symmetric than the Lockwood array, furthermore the size of the two halves of the polar plot are much closer to the individual microphone.

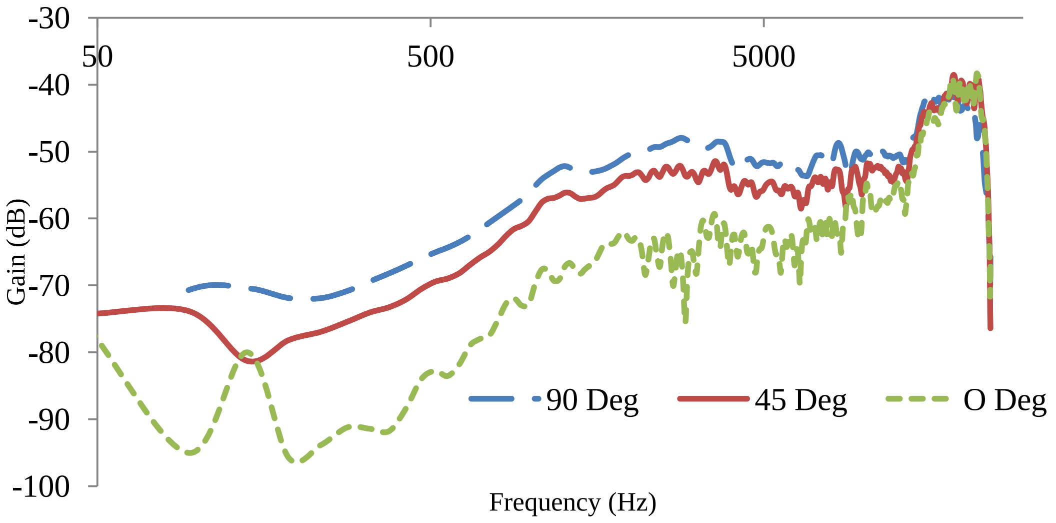


Figure 27 : Frequency Response of the pressure gradient microphone in the AVS II array.

This result show that when the microphones are placed adjacent to each other errors are introduced in the directional plots due to the increase in surface area and the increase in separation between the front and the back of the microphone. Hence, when the microphones are moved these errors are reduced. But even with the offset at higher frequencies, the symmetry of the plots is not exactly correct.

This change in design has provided an improvement, but there is one area of the design that could be improved such that further improvements in performance can be achieved. In this design, the support that holds that microphones in place is almost as wide as a microphone capsule, hence this square pole will also introduce error in the

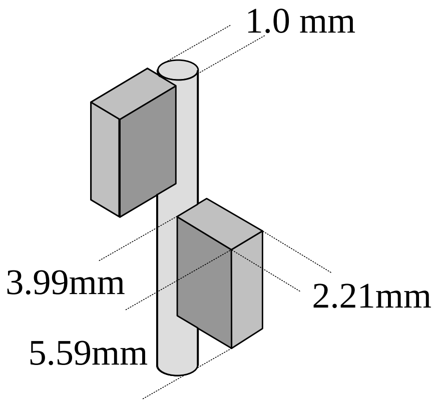


Figure 28 : Dimensions of AVS III.

directional response of the microphone at higher frequencies. By reducing the size of the pole further improvements in the directional response can be achieved. The array presented in the next section is the final improvement that is proposed to the Lockwood array, the term that will be used in describing this array will be AVS III.

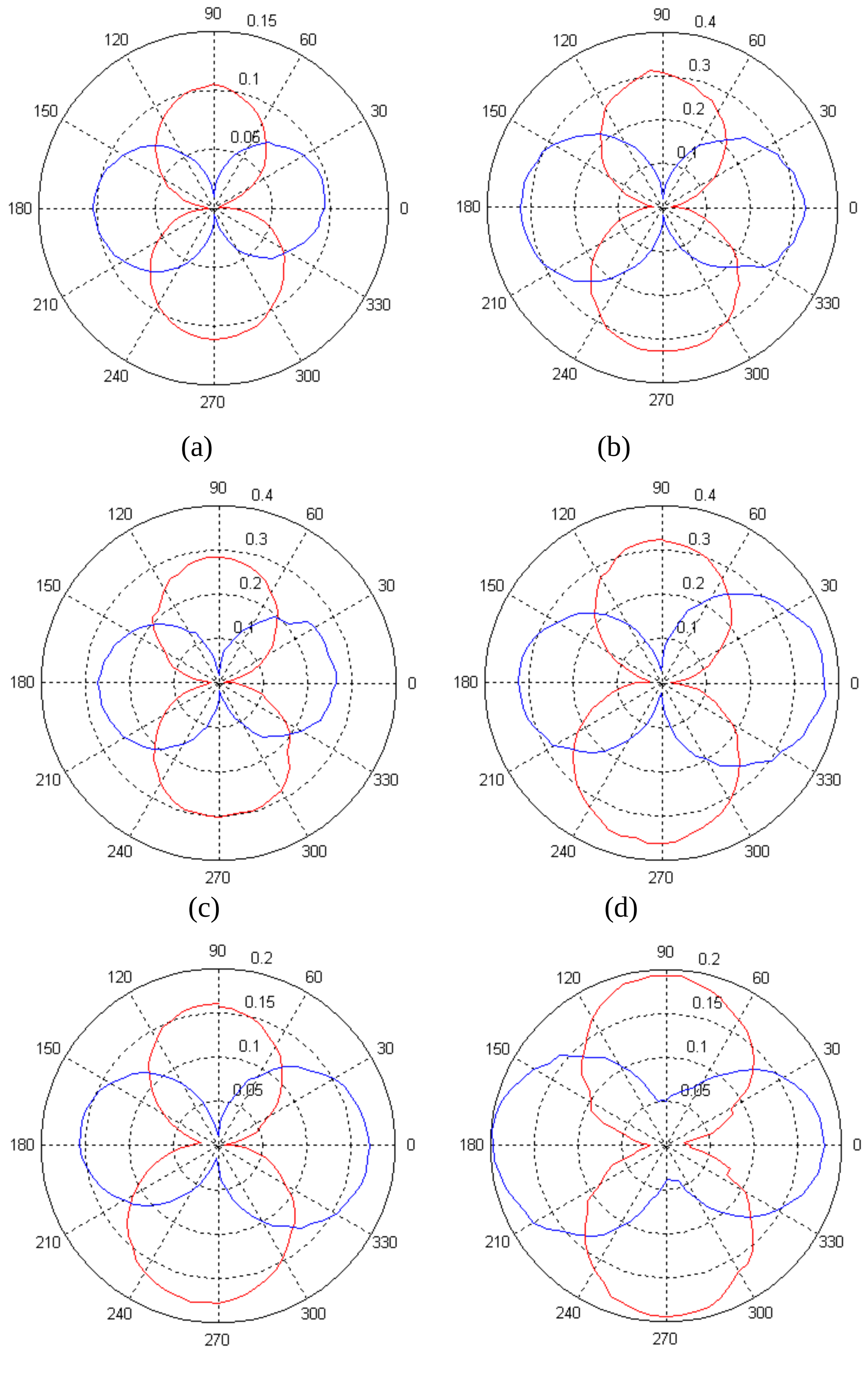
3.2.11 The AVS III array

The AVS array presented in the previous section showed good improvement in the directional response. Here, further improvement to the AVS II is achieved by reducing the size of the support holding the microphone capsules in place.

The light weight and small size of the microphone capsules does not require a large metal support to hold the microphones in place, rather a thin metal rod which is capable of holding the microphones in place is enough. A 1mm diameter metal rod is chosen to hold the microphone in place, the dimensions of the new array is shown in Figure 28. There are two advantages that are offered by the metal rod, which are:

1) Due to the small diameter of the metal rod support does not contribute to an increase in distance separating the front to back of the microphone and
2) It does not contribute to an increase in surface area.

Unlike the square aluminium pole which has an edge on one side of the microphone, the metal rod does not have any sharp edges that could contribute to reflection or diffractions.



(a) (b)

(c) (d)

(e) (f)

Figure 29 : The Directional Response of pressure gradient microphones on AVS II array. The x sensor is plotted in red and y sensor in blue (a) 100 Hz (b) 500 Hz (c) 1 kHz (d) 3 kHz (e) 5 kHz (f) 7 kHz.

The only complication in this design is placing the microphones exactly at 90 degrees to each other. Hence, to hold the microphones in place a special rig was produced which aligns the microphones in place for attachment. The AVS III array is shown in Figure 30, with the small metal rod holding the microphones in place.

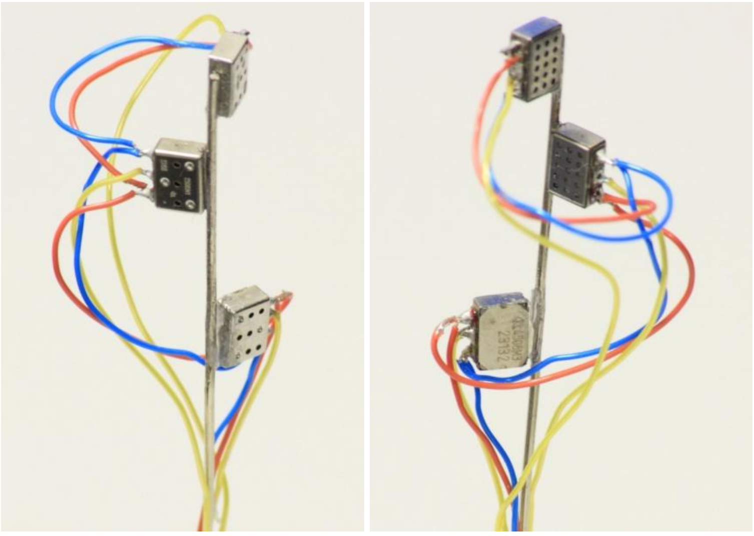

(a) (b)

Figure 30 : The AVS III (a) front showing the *x* and *y* sensors (b) back showing the Omni-directional sensor.

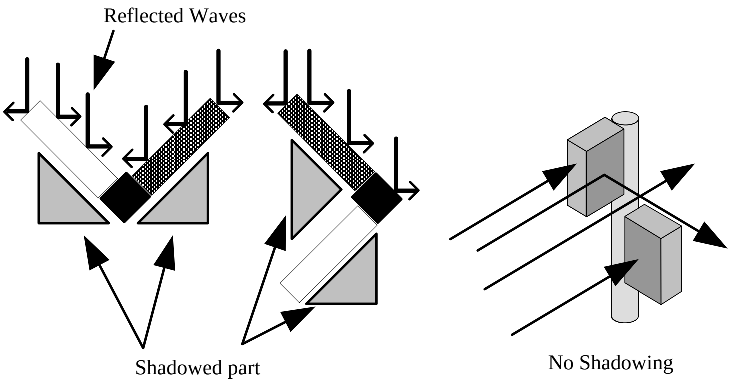


Figure 31: The effects of shadowing and reflection in the Lockwood array.

3.2.12 The Frequency and Directional Response of AVS III

The frequency response of the microphones attached to the AVS array is examined as outlined in previous sections. The result of the frequency response for the

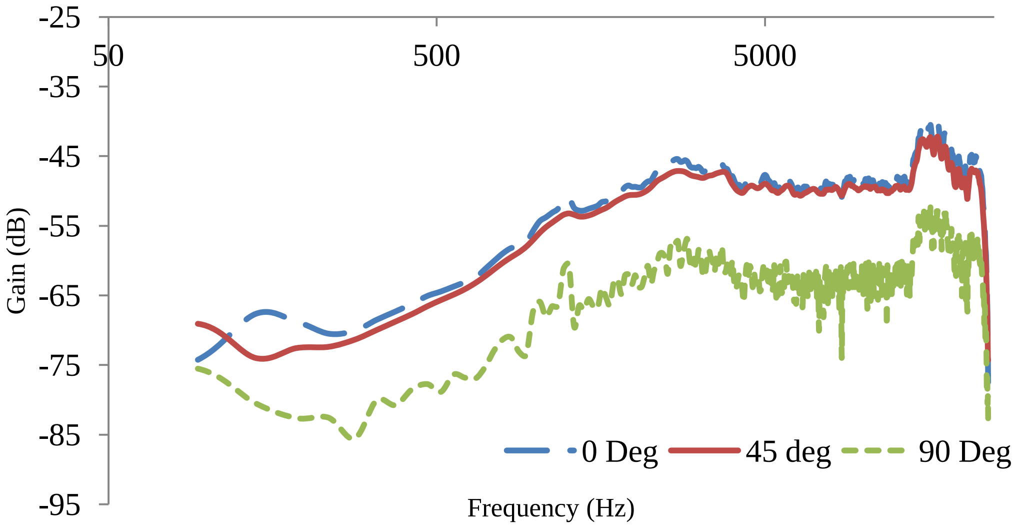


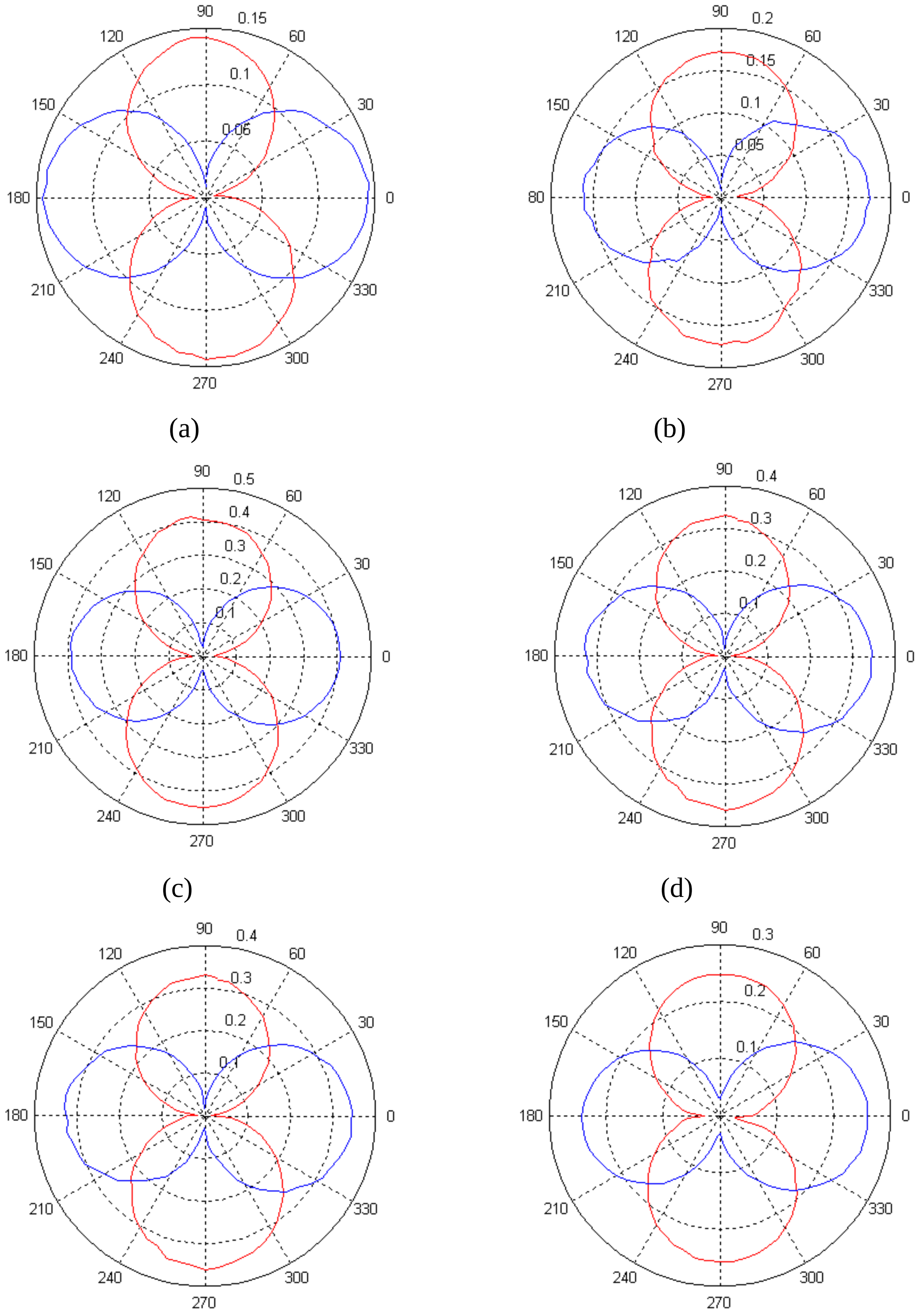
Figure 32 : Frequency Response of the pressure gradient microphone in the AVS III array.

pressure gradient microphone in the AVS III array is shown in Figure 32. The result for the frequency response of the AVS III is very close to the frequency response of the individual pressure gradient microphone. This is expected as it can be seen from Figure 28 and Figure 30 the microphones on the array are virtually without any obstructions from the support.

The directional response of the AVS array is presented in Figure 33, for selected frequencies between 100 Hz and 10 kHz. From the plots of the directional responses, it is seen that at all frequencies, the symmetry of the figure-of-eight plots are maintained. Furthermore the plots are smoother than that of the Lockwood array and AVS II.

The effect of shadowing for the Lockwood array was discussed in Section 3.2.7, here, with the AVS III it can be seen from Figure 31 by offsetting the directional sensors, the effect of shadowing is completely removed, and the errors due to the effects of shadowing, reflection, and diffraction minimised in the AVS III.

There are some errors in the plots shown in Figure 33 such as small imperfections in the smoothness of the polar plots and mid part of figure-of-eight at higher frequencies are wider. These errors are due to the imperfections in attaching the microphones on the support and due to wires connected to the microphones, How ever since these errors are very small they were found to be tolerated in the applications described in this thesis.



(a) (b)

(c) (d)

(e) (f)

Figure 33 : The Directional Response of pressure gradient microphones on AVS III array. The x sensor is plotted in red and y sensor in blue (a) 100 Hz (b) 500 Hz (c) 1 kHz (d) 3 kHz (e) 5 kHz (f) 7 kHz.

3.3 The Output of AVS Array

The output of AVS consists of two components: an acoustic particle velocity and acoustic pressure component. This can be expressed in vector form as:

$$\mathbf{y}(t) = [o(t) \quad x(t) \quad y(t) \quad z(t)] \tag{113}$$

where $o(t)$ represents the acoustic pressure component and $x(t), y(t)$ and $z(t)$ represents the pressure gradient components. The relationship between the acoustic pressure and the particle velocity is given in (4), and the output from a pressure gradient microphone is given in (112). This is true for a single pressure gradient microphone, but for the AVS array as whole the relation between the particle velocity and acoustic pressure for all the array elements can be expressed in terms of the steering vector as:

$$[x(t) \quad y(t) \quad z(t)] = \left[\frac{j}{k_0 z_0}\nabla p\right]\mathbf{u} \tag{114}$$

where **u** is the steering vector for an AVS array, which is expressed as:

$$\mathbf{u} = [1 \quad cos\theta cos\emptyset \quad sin\theta cos\emptyset \quad cos\emptyset] \tag{115}$$

where θ is the azimuth and angle and $\emptyset$ is the elevation angle. The general form of the signals at the output of the AVS in both anechoic and reverberant conditions can be expressed as:

$$x(t) = g_x cos\theta s(t) + h_x s(t) + g_x n(t) \tag{116}$$

$$y(t) = g_y cos\theta s(t) + h_y s(t) + g_y n(t) \tag{117}$$

$$z(t) = g_z cos\theta s(t) + h_z s(t) + g_z n(t) \tag{118}$$

$$o(t) = g_o s(t) + h_o s(t) + g_o n(t) \tag{119}$$

where $n(t)$ represents diffuse noise and $s(t)$ is the source signal at an angle θ, to the microphone and terms g_x, g_y and g_o represent the gains of the microphones (these may differ due to mismatches in capsule responses and inaccuracies due to AVS array construction). Filters h_x, h_y, h_z and h_o model the multipath effects of the source signal to the microphone as well as mismatches between microphone capsule responses and inaccuracies due to AVS array construction. In anechoic conditions the multipath effects and noise terms are 0. The array used in this work is a two dimensional array hence only the o, x and y components will be used.

3.4 DOA Estimation for Measuring the Performance of the Array Design

There are several methods for estimating DOA for an AVS, one method for estimating DOAs for an AVS array is the MUSIC algorithm, which was discussed in detail in Section 2.9.9. The MUSIC algorithm has been used with velocity hydrophones to estimate DOAs in [130, 131] where the application is based on random array configurations and simulations showed accurate estimates of the DOA. The MUSIC algorithm allows for the estimation of the DOA using the Eigen-values and Eigen-vectors of the covariance matrix formed from the recorded signals. The MUSIC algorithm is given in (61). Since the array used in here is a two dimensional array and all the sources are at 0 degrees in elevation, the steering vector for the MUSIC algorithm in (61) is expressed as:

$$\mathbf{u} = [1 \quad cos\theta \quad sin\theta] \tag{120}$$

The other methods for estimating DOA for an AVS are based on the ratio of the intensities of the x and the y components. Although these algorithms provide accurate DOA estimates, here the purpose of obtaining DOA estimates is to evaluate the design of the array; hence a reliable and well known approach that can be used with any array configuration which has been proven to give accurate DOA estimates is more convincing than an approach that is unique to an AVS. Hence, the MUSIC algorithm is used for evaluating the performance of the AVS arrays.

3.4.1 Array Calibration for DOA Estimation

The microphone capsules used in the construction of the AVS arrays are all commercially available microphones, which is designed with a built in Field Effect Transistor (FET) amplifier. Due to the internal circuitry it was found that the microphones do not always produce the same output levels for a constant test source and seen from Figure 25, 28 and 33 where the gain of the x and y microphones are different. To estimate the DOA estimation from the array, the ratios between the x and y components has to be according to (120). Hence, it is important that the output levels of the microphones be calibrated before recording. To compensate for these errors, a gain correction factor was determined through analysing recordings from three separate AVS

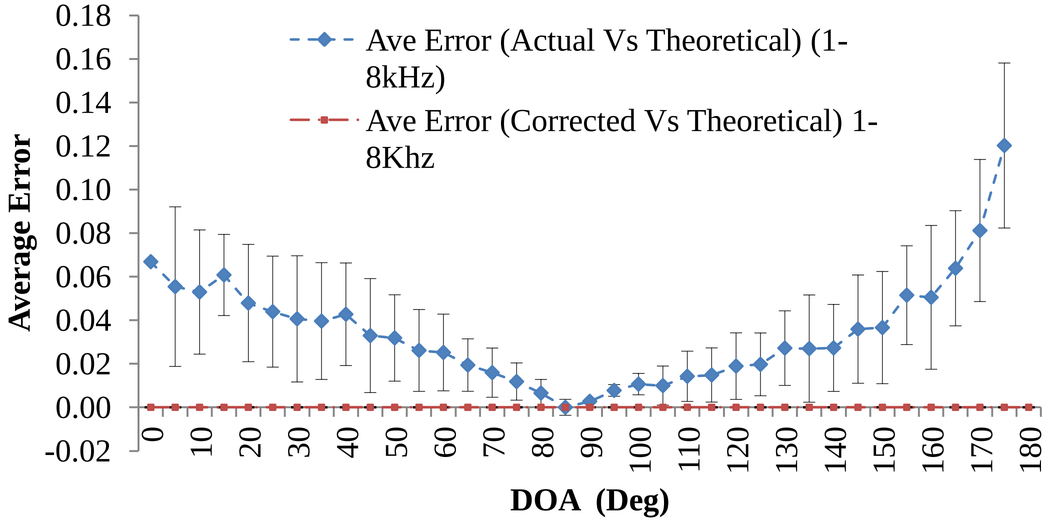


Figure 34 : Average Error between the actual Vs theoretical and Corrected Vs Theoretical for 1 kHz-8 kHz monotone signal (Error bars indicate 95% confidence intervals).

arrays of a series of monotones ranging in frequency from 1 kHz to 8 kHz (in steps of 1 kHz) and for directions ranging from 0 degrees to 360 degrees (in steps of 5 degrees).

The gain levels on the preamplifiers are set to the same level by adjusting the gain levels by placing the x microphone is exactly at zero degrees to the source and as a test tone is recorded the gain is recorded and the array is then rotated such that the y microphone is exactly at zero the gain of the y channel is adjusted until the output is exactly the same as the x channel.

The theoretical values for each direction are found for each channel from the maximum energy value which is equal to both channels. A simple correction method consisting of the average ratio of actual to theoretical polar response at each direction is determined. Figure 34 shows the resulting polar response error (the difference of the recorded to theoretical response) as a function of source direction. Compared with the non-compensated recordings, the compensated recordings have significantly less error and are statistically equivalent to the theoretical response as measured by 95 % confidence intervals.

3.4.2 Localization Experiments

Results for localization were obtained using the same experimental rig, recording environment and sound sources described in Section 3.2.3. As well as the three AVS configurations, recordings were also made with a four element ULA, which was chosen so that the number of sensors matched the AVS. The ULA was built using the same Knowles EK-3132 omni-directional microphones as used in the AVS and using a spacing of 21 mm; this results in an array of approximately 42 mm long. Localization was performed using the DOA estimate described in the previous section for frequencies 1 kHz to 10 kHz in 1 kHz intervals and for the all four quadrants i.e. $\theta \in (0,2\pi)$.

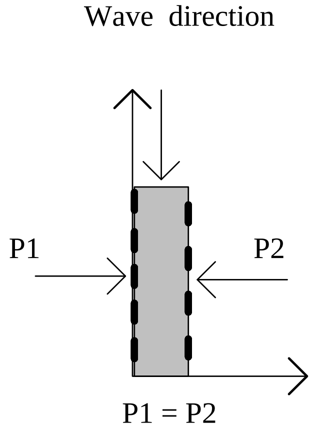


Figure 35: Pressure distribution at 0 degrees around the pressure gradient microphone.

3.4.3 Response to Source Perpendicular to Sensor Inlets

An ideal gradient microphone should have an output of 0 for sources located perpendicular to the sensor inlets; this is because the pressure will be identical on either side of the microphone as illustrated in Figure 35 and hence the pressure gradient (or difference) should be zero.

To analyse this characteristic, the Average Angular Error (AAE) of signals impinging on the AVS at 0 degrees was measured for the Lockwood Array, AVS II and AVS III for frequencies from 1 kHz to 10 kHz using:

$$AAE = \frac{1}{N}\sum_{n=1}^{N}\left|\theta_{n,m} - \theta_{n,a}\right| \qquad (121)$$

where N is number of sources (tones) and $\theta_{n,m}$ and $\theta_{n,a}$ are the measured (m) and actual (a) DOAs, respectively, for source n. Figure 36 shows the difference error for sources located at 0 degree to the Y axis; the error for Lockwood array is much higher than that compared to AVS II with offset sensors and AVS III. Furthermore, the results

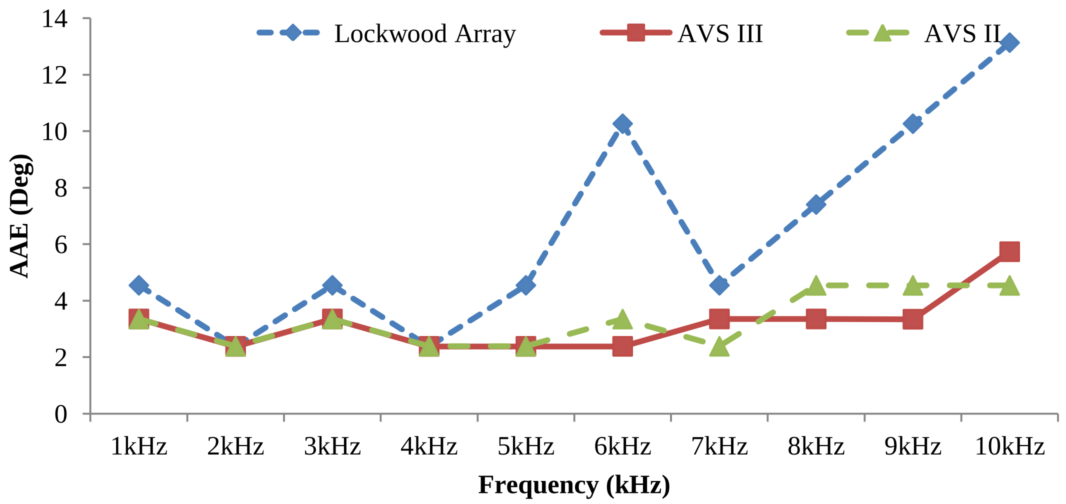


Figure 36: The AAE for output at 0^0 for frequencies 1 kHz to 10 kHz

show that that the difference in error increases as the frequency of the source signal increases.

In the case of the Lockwood array when the source is at 0 degrees to the y sensor, air particles flowing on the x sensor side see a larger separation between the front and the back of the microphone and also a larger surface area which contributes the increases in errors. Furthermore, at higher frequencies the x sensor and support cause reflection and diffraction of the soundwave hence a slight increase in pressure on one side of the sensor; this increase in pressure causes errors in the output. When the sensors are placed at an offset as in AVS II the error is reduced significantly and for frequencies up to 5 kHz the error is the same as that for AVS III. The improved result is due to the offsetting of the sensors which reduces the effects of reflection and diffraction and reduces the separation between the front and the back of the microphone as well as the surface area, hence creating an output which is more accurate than the output of the Lockwood array.

3.4.4 Direction of Arrival for Lockwood Array, AVS II and AVS III

The average angular error for sources located in the 1st quadrant and averaged over all recorded source frequencies is shown in Figure 37. On average, the AAE is 1.5 degrees for AVS III, for AVS II AAE is 3.2 degree and 4.6 degree for Lockwood array. The results of Figure 37 are for the AAE of the DOA estimates for the second quadrant. On average, the AAE for AVS III is 1.9 degree while for AVS II is 9.5 degree and for

Lockwood array is 7.3 degree. Compared to the results from the 1st quadrant the overall error for all AVS's is seen to have increased significantly for the second quadrant. The results for second quadrant are for frequencies 1 kHz to 6 kHz as the error from frequencies including 7 kHz and above were statistically not reliable for Lockwood array.

For the first quadrant it is proposed that the increase in error for Lockwood array is caused by the artificial increase in the front and back separation, surface area and the effects of reflection, diffraction and the acoustic shadowing at high frequencies. A significant improvement in error is seen when the sensors are at offset (see AVS II results and AVS III). It is proposed that this improvement is due to reduced blocking from any object to the flow of the air particles; hence the sensor readings are more accurate.

For the second quadrant, the effect of the artificial increase in front to back separation and the surface area is more than that for the first quadrant. In addition to this, the reflection from the edges of the square pole at high frequencies and shadowing or blocking by the sensor on the opposite axis at high frequencies contribute more when the source is positioned towards the back of the array as is in quadrant 2. It is believed that the reflections and blocking from the square pole has a greater significance on the error as suggested by the results in Figure 38. By removing the square pole and replacing it with the thin metal pole there is a significant reduction in error for AVS III. In Figure 31, it can be seen that the impinging soundwave hits the sensor and the square pole and the waves are reflected creating regions of attenuations or shadowed parts. These cause incorrect readings of pressure difference that produce an error in the output for Lockwood array, and for the AVS II the square aluminium pole cause reflections which create errors in the output. In contrast, for AVS III the metal pole is much smaller than the sensors, which are also offset; this results in a pressure difference and an output with minimum error.

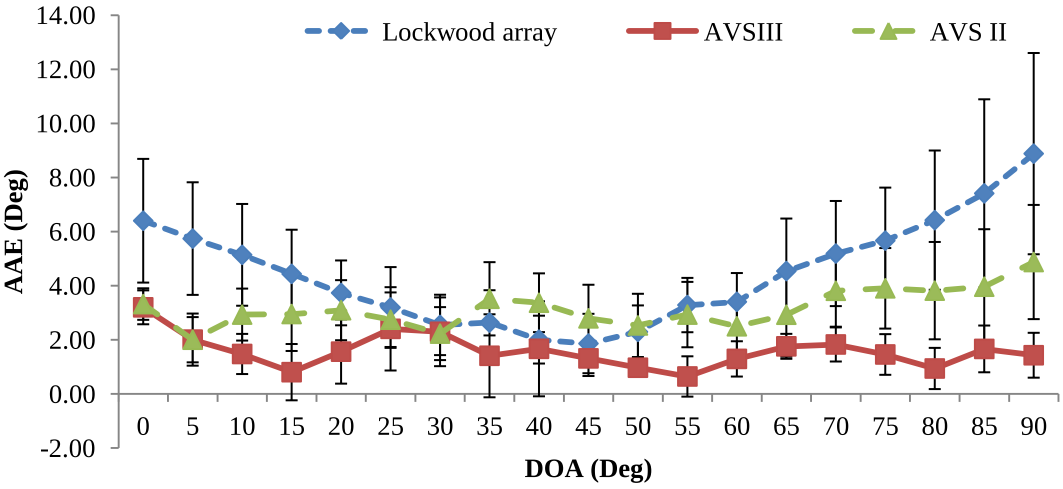


Figure 37 : AAE of the DOA estimates for 1st quadrant. Error bars represent 95% confidence intervals.

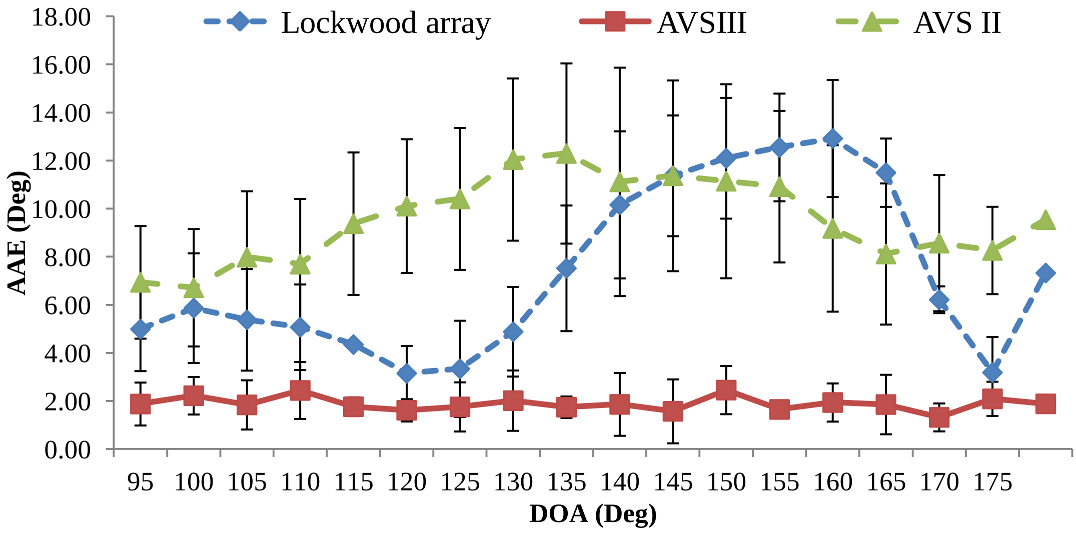


Figure 38 : AAE of the DOA estimates for 2nd quadrant. Error bars represent 95% confidence intervals.

The DOA estimates for the third and fourth quadrant are shown in Figure 39 and Figure 40 where when the position of the source is behind the array the error for the Lockwood array increase sharply, where as for the arrays with the sensors offset the errors remain low. Hence, from these results it can be said when source is at the back of the array the artificial increase in front to back separation and the surface area is maximum and hence the errors are at a maximum. Furthermore, the error bars for the Lockwood array in the third and forth quadrant is much larger than that of the first and

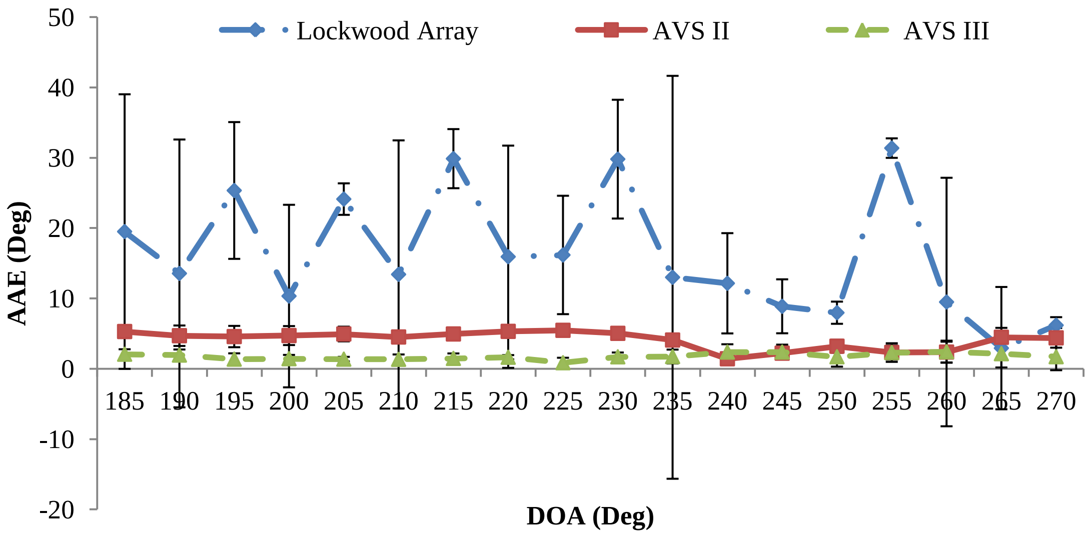


Figure 39 : AAE of the DOA estimates for 3rd quadrant. Error bars represent 95% confidence intervals.

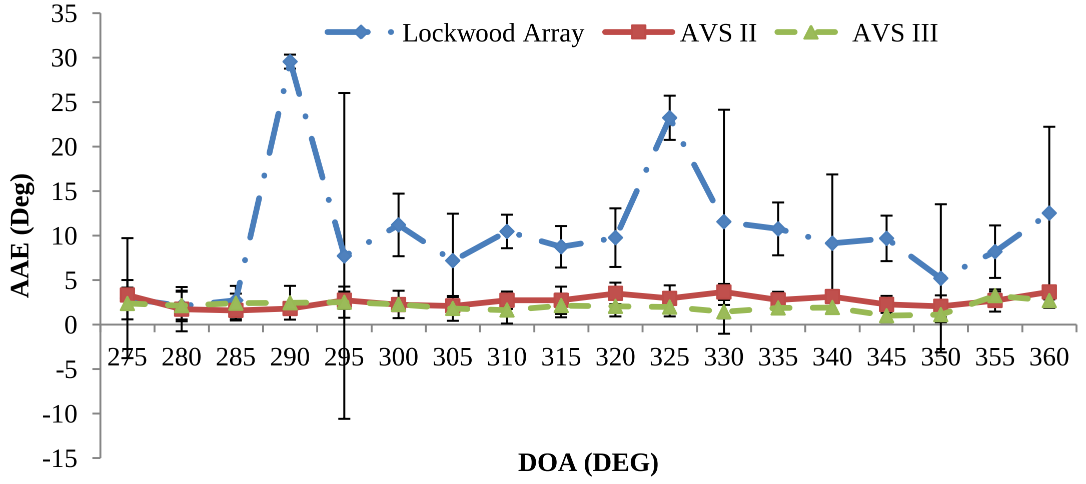


Figure 40 : AAE of the DOA estimates for 4th quadrant. Error bars represent 95% confidence intervals.

second quadrant, and due to the larger errors bars it can be said that the results from the Lockwood array are statistically invalid.

3.4.5 DOA Estimates Vs Frequency for AVS

Results in Figure 41 show that as the frequency of the source increases the error in the DOA estimate also increases. The average AAEs for source frequencies from 1 kHz to 10 kHz are approximately: 6.1 degree for Lockwood array, 5.0 degree for AVS II; and 2.4 degree for AVS III. For the Lockwood array, the AAE versus source frequency is

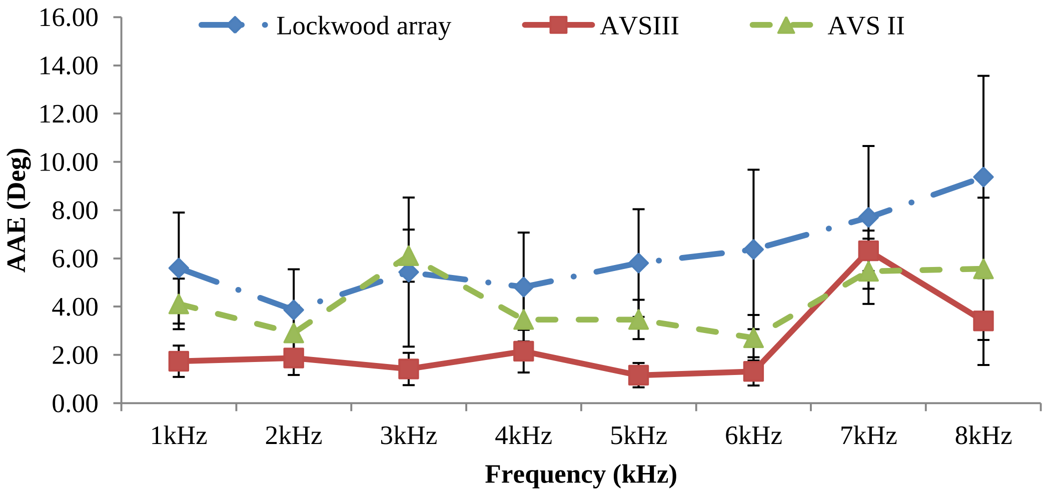


Figure 41 : AAE for each frequency Band vs Frequency. Error bars represent 95% confidence intervals (top half of error bar for 10 kHz removed for clarity).

approximately constant up to 5 kHz, increases by approximately 2 degree per kHz between 6 and 9 kHz before increasing sharply to 20 degree at 10 kHz. The AAEs for AVS II remain below 9 degrees for source frequencies up to 10 kHz, while for AVS III the maximum error is 8 degree (except at 9 kHz).

This result shows that by offsetting the sensors and reducing the surface area of the structure holding the AVS microphones, more consistent and accurate DOA estimates can be obtained for all source frequencies tested. By first offsetting the sensors, a reduced DOA estimate error is achieved for high frequencies (as seen by the results for AVS II). Replacing the square pole with a cylindrical pole of much smaller area leads to further reductions in the DOA estimate errors for all frequencies.

3.4.6 Comparison of DOA Estimates for AVS and ULA

The ULA is the simplest and most common type of microphone array, which has been described in detail in Section 2.8.3. For optimum performance of a ULA the spacing between the microphones has to be set logarithmically, but since they are only 4 microphones they have been attached with the same separation as shown in Figure 42. In this experiment, only three of the four available microphones on the ULA is used, this is done in order to make the comparison with the AVS valid. The results in Figure 43 are a comparison of the DOA error produced by a ULA to that of the AVS. The

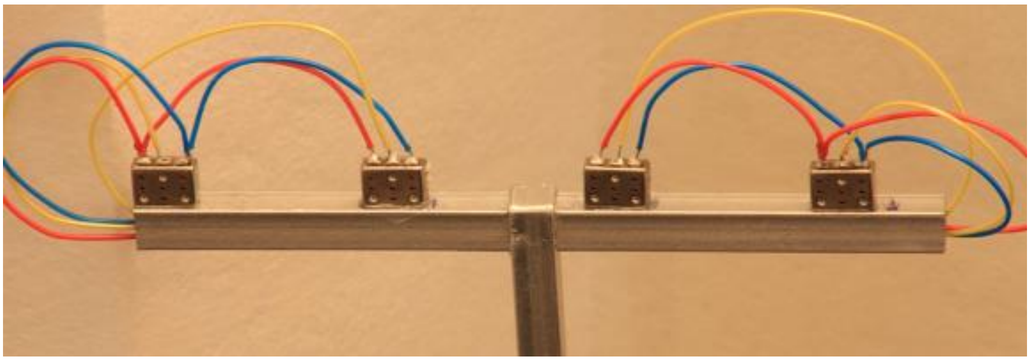


Figure 42 : A four element ULA array – the microphone capsules used in the array are Knowles EK 3132 omni-directional microphones.

steering vector (also known as the array response vector) for the ULA used for obtaining the MUSIC spectrum is given in (37).

The results show that at a distance of only 1 m from the source, AVS III has an average error of 1.6 degree compared to that of ULA which has an average error of 21.8 degrees. The Lockwood array has an average error of 4.5 degree which is the worst result for all AVS's but is still approximately 4 times better than the average error produced from a ULA with the same number of microphones and comparable size. This results show that the performance of AVS's are much better than ULA's of comparable size.

To produce results comparable to the AVS, the ULA would need to be placed much further (at least 2.5 m to 3 m) from the source and use more microphones (at least 5) with much larger separations [26] as explained in Chapter 2. Preferably, microphones should be separated logarithmically so that the array responds accurately to tones of different frequencies [26].

3.5 Conclusions and Summary

The work presented in this chapter has shown an AVS design that delivers highly accurate estimates of DOA for in-air applications. The results obtained show that there is significant impact on the directional response and the DOA estimates by:

- The artificial increase in distance separating the front and back of the microphone due to the adjacent microphone and the structure holding microphones in place.

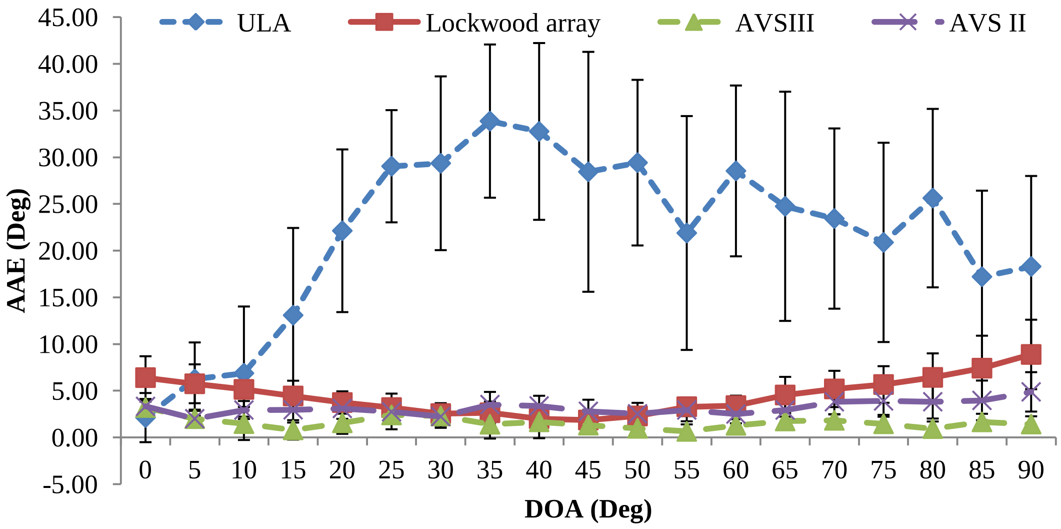


Figure 43: AAE for DOA estimates for AVS I, AVS II and ULA. Error bars represent 95% confidence intervals.

- The artificial increase in surface area of the microphone due to the structure and the adjacent microphone.
- Acoustic shadowing, reflection, and diffraction due to the adjacent microphone and the structure holding the array together at higher frequencies.

The results show that by placing the sensor such that the sensor on the off axis does not block the path of the adjacent sensor the result of the directional response and DOA estimates are improved significantly for the all quadrants.

Furthermore, the results show that by changing the shape of the support from square to cylindrical, which reduces the cross sectional area of the support by 5.46 mm^2, provides a significant improvement in the estimated DOA accuracy. The DOA estimates obtained from the new design have an average error of less than 2 degree for a range of source frequencies, compared with average errors of more than 4.5 degree for an alternative existing design. Furthermore, it has been established that the accuracy of the DOA estimates generated by the AVS is much better than the estimates for a ULA with similar number of sensors and comparable size at close proximity to the target source. The next chapter will examine applications of AVS for DOA estimation in reverberant conditions for monotone and speech signals.

Chapter 4 DOA Estimation for an AVS

4.1 Introduction

The most important information from any microphone array is the DOA estimate of the desired source. The DOA estimate is vital for other algorithms such as beamforming, source separation and dereverberation. The target applications of the AVS array are hands free communications and teleconferencing with mobile devices. Hence, the ability to locate sources without a large microphone array would be extremely useful for such devices. For applications such as mobile hands free teleconferencing, this feature would enable to focus on a given source to capture, steer a camera towards the source and enhanced recordings with ease.

In Chapter 3, the AVS design was considered with a solution presented that resulted in the AVS being capable of producing accurate DOA estimates of mono-tone stationary sources with errors of less than two degrees in anechoic conditions, but in real life applications it is very rare to find perfect anechoic conditions. Hence, it is important to evaluate the performance of DOA estimation with an AVS in reverberant conditions with background noise and for real sources such as speech. Furthermore, in real conditions, the sources may not be stationary and there may be more than one source. Hence, it is vital to evaluate the performance for moving sources and multiple sources. In this chapter DOA estimation will be performed on recordings made under reverberant conditions with considerable background noise and for stationary, moving and multiple sources.

One of the most complicated problems in DOA estimation is to obtain the direction of the sources when multiple sources are present and the sources overlap. Here, DOA estimation for one, two and three sources will be presented with a comparison between the performances of different algorithms for DOA estimation.

The target applications for the DOA estimation with an AVS are real time applications, which require real time processing of the data with minimum delay and eventual transmission over a telecommunications channel. Hence, it is vital that DOA estimates be made in real time with minimum delay. In traditional approaches, to get an accurate estimate of the source direction, multiple frames of recorded signals are required, which introduce delays into the system. Here, it will be shown that with the

AVS a single 10ms frame is enough to get an accurate estimate of the source direction for stationary, mobile, and multiple sources.

The only microphone array that closely resembles an AVS in terms of how the signals are captured is the Soundfield Microphone which has four cardioid pressure sensors arranged in a tetrahedron configuration as described in Section 2.8.7. Unlike the AVS, the Soundfield produces the W,X, Y and Z directional components by combining the four capsule signals. Here, results are compared for DOA estimation using both the AVS and Soundfield microphones.

Most work done on DOA estimation and speaker tracking is based on the Time Delay Estimate (TDE) or TDOA with non co-incidental microphone arrays. In [132] six pairs of four microphones are used to track and find DOA estimates using non-linear particle filtering. In [133] three Soundfield Microphones are positioned in a straight line to form a microphone array with known geometry and using the X and W components only source localization is achieved. In [134] binaural microphones are used to track multiple speakers in a cocktail party situation.

In reverberant environments, these TDE based approaches are less accurate due to sound reflections. In contrast, since microphones are co-located, the AVS does not rely on TDE for source localisation estimation. Here, the MUSIC algorithm and intensity based algorithms for DOA estimation will be used. Due to the use of highly directional sensors, the AVS provides many advantages over other microphone arrays for DOA estimation. In particular, the secondary reflections in reverberant conditions are minimised due to two features of the array,

a) The co-location of the sensors,

b) The directionality of the sensors.

There are post-processing techniques for improving the localisation accuracy for spaced microphone arrays [134-136]. However, in this work, the focus is on investigating the advantages that can be drawn from the AVS without such post-processing techniques. The motivation is to minimise additional computational complexity for use in real time applications such as video teleconferencing. The work presented in this chapter is unique as this is the first time a single co-located microphone array is used for DOA estimation of speech sources in reverberant conditions, for moving sources and for multiple sources.

The remainder of this chapter will be organised as follows: Section 4.2 will present the different types of DOA estimation algorithms that can be used with an AVS. Section 4.3 will present an outline on the experimental setup and the database of the speech used in the experiments. Section 4.4 will outline the results for experiments for single stationary and moving sources and Section 4.5 will present the results for multiple sources and finally Section 4.6 will give a summary of the results presented in this chapter.

4.2 DOA Estimation in the Time and Frequency Domain Using an AVS

The pressure gradient sensors of the AVS capture the sound pressure as well as directional information, which can be used in the calculation of the DOA estimates from array outputs. The steering vector in (115) is a combination of (31 to 33) in a single vector, which give the position of the source in three dimensional space. From (112) it can be seen that the output of the microphone is a direct representation of particle velocity. The particle velocity as a vector in the *x* and *y* directions can be expressed as:

$$\mathbf{v_x} = \frac{j}{k_0\rho_0c_0}[\nabla\, p_x\mathbf{u}_x] \tag{122}$$

$$\mathbf{v_y} = \frac{j}{k_0\rho_0c_0}[\nabla\, p_y\mathbf{u}_y] \tag{123}$$

where $\mathbf{u}_x$ and $\mathbf{u}_y$ are the unit vectors in the *x* and *y* directions and $\mathbf{v}_x$ and $\mathbf{v}_y$ are the unit vectors of particle velocity in the x and y direction. The Instantaneous intensity of a soundwave is expressed as the product of the sound pressure and the particle velocity [20, 21].

The instantaneous intensity at a point due to a soundwave is the product of the particle velocity of that wave and the pressure. Hence, the instantaneous intensity due to the x and y components can be expressed as:

$$\mathbf{I_x} = \frac{j}{k_0\rho_0c_0}[\nabla\, p_x\mathbf{u}_x]P \tag{124}$$

$$\mathbf{I_y} = \frac{j}{k_0\rho_0c_0}[\nabla\, p_y\mathbf{u}_y]P \tag{125}$$

where $\mathbf{I}_x$ and $\mathbf{I}_y$ are the instantaneous intensities in the x and y directions and P is the sound pressure. The output at the omni-directional microphone is a direct representation of the sound pressure at the AVS similar to (112).

Based on (124) and (125) $\mathbf{u}$ can be estimated by phasor time averaging and renormalization [7]. The time domain estimate of the source direction can be found from:

$$\hat{s}(1,1) = \frac{1}{N}\sum_{t=1}^{N} Real\{o(t).x(t)\} \tag{126}$$

$$\hat{s}(1,2) = \frac{1}{N}\sum_{t=1}^{N} Real\{o(t).y(t)\} \tag{127}$$

the estimate of $\mathbf{u}$ is calculated as [7]:

$$\hat{\mathbf{u}} = \frac{\hat{s}}{||\hat{s}||} \tag{128}$$

where $||\hat{s}||$ is the Euclidian norm of $\hat{s}$ and $Real$ is the real part of $\{o(t).x(t)\}$. Here, the DOA estimates are found on a frame-by-frame basis and hence this method should in theory work on both monotone and complex signals such as music and speech.

Since most signals in practice are complex signals with a broad range of frequencies, DOA estimation in the frequency domain can give advantages over time domain implementations. From (124) and (125) the direction of the instantaneous intensity can be expressed in the frequency domain as [137, 138]:

$$\theta(k) = \tan^{-1}\frac{\mathrm{I}(k)_y}{\mathrm{I}(k)_x} \tag{129}$$

here k is the discrete frequency and $\mathbf{I}(k)_x$ and $\mathbf{I}(k)_y$ are calculated by applying an FFT to signals (116), (117) and (119). The resulting direction is obtained as follows [137, 138]:

$$\theta(k) = \tan^{-1}\left[\frac{Real\{O(k)^{*}\times Y(k)\}}{Real\{O(k)^{*}\times X(k)\}}\right] \tag{130}$$

where $Real$ is the real part of the FFT of the channel and $*$ is the conjugate. The directions calculated from (130) are for each frequency component of the current frame. The advantage of calculating the directions for each frequency component is if there are two or more sources with different frequency components, then the directional information for each source can be useful in separating the sources.

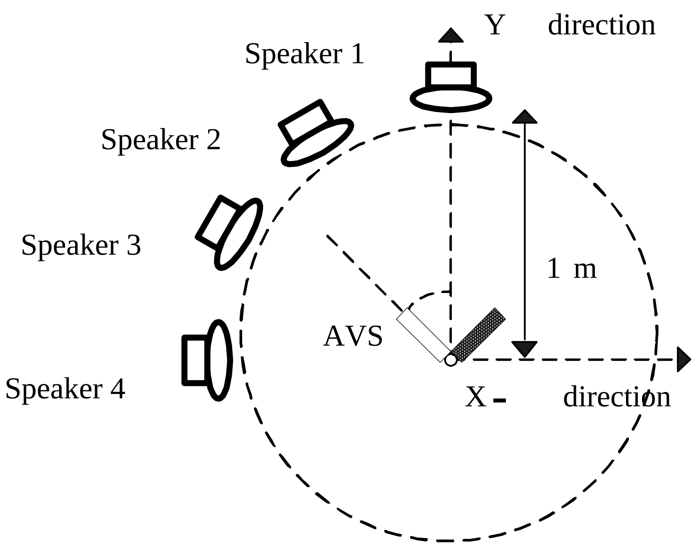


Figure 44 : Experimental setup for DOA estimation of single, multi and moving sources

4.3 Localization Experiments

4.3.1 Experimental Setup

Recordings were made in a reverberant room with RT_{60} of 30ms and with considerable background noise of computer servers and air-conditioning at 53.1dBA. For testing, the experimental setup of Figure 44 was used, where the AVS was mounted on a custom built rotating platform (to allow positioning of the microphones relative to the source) and self powered loudspeakers (Genelec 8020A) was placed in front of the AVS at a distance of 1m with an elevation of 0 degrees. A series of monotone signals each two seconds long and of equal energy were played with frequencies ranging from 1 kHz to 10 kHz. For speech, five male and five female sentences from the IEEE speech corpus [139], each approximately two and a half seconds long with different speeds were played. Recordings were made at 5 degree intervals, with a sampling rate of 48 kHz, which for the case of the frequency domain DOA estimation algorithms were down-sampled to 16 kHz.

The multi-source recordings were made for two sources and three sources. For two sources loud speaker 1 was kept stationary and loud speaker 4 was moved in increments of 15 degrees from 0 to 90 degrees. For three sources loud speaker 1 and 4 were kept stationary and loud speaker 3 was moved from 15 degrees to 75 degrees in increments of 15 degrees.

The results present in this work are for average angular error which is the error between the actual angle and the angle obtained from the DOA estimate, which is calculated according to (121). The results presented in the following sections are for confidence intervals of 95 %.

4.3.2 Monotone Stationary Sources

The first experiment performed in this section is to calculate the DOA estimates from different algorithm using an AVS and a Soundfield microphone for monotone signals. Three different DOA estimation algorithms are analysed here, which are: MUSIC algorithm; time domain intensity algorithm for DOA estimation; and frequency domain version of the intensity algorithm for DOA estimation.

In the previous chapter the results for the MUSIC algorithm with the AVS for monotone signals in anechoic conditions showed very accurate results. Hence, here the MUSIC algorithm can be used as a benchmark of the other two algorithms.

The signals are processed on a frame by frame basis with an overlap of 50% and a frame length of 20ms. The FFT length for the frequency domain implementation is set at 512 points. The output of the frequency domain implementations results in N DOA estimates per frame, where N is the length of FFT hence the average of all the DOAs for each frame is used. For the test where only one source is involved, the results shown are for average DOA for all frames analysed for the time domain algorithm and for the frequency domain algorithm the results presented are for the average of all the DOAs for each frame and averaged over all frames.

Figure 45 shows the results for AAE for monotone signals over a rotation of 90 degrees in azimuth at 5 degree intervals for the AVS. The results show that the AVS has an average error of 0.98 degrees for the MUSIC algorithm while the average error for intensity based algorithms for AVS are 1.64 and 1.68 degrees for time domain implementation and frequency domain implementation, respectively. This result shows that the MUSIC algorithm performs better than the intensity based algorithm.

The intensity based algorithms require the microphone gains to be exactly the same when recordings are made, but this is not practically possible especially in reverberant conditions. In these experiments the gains of the microphones are adjusted

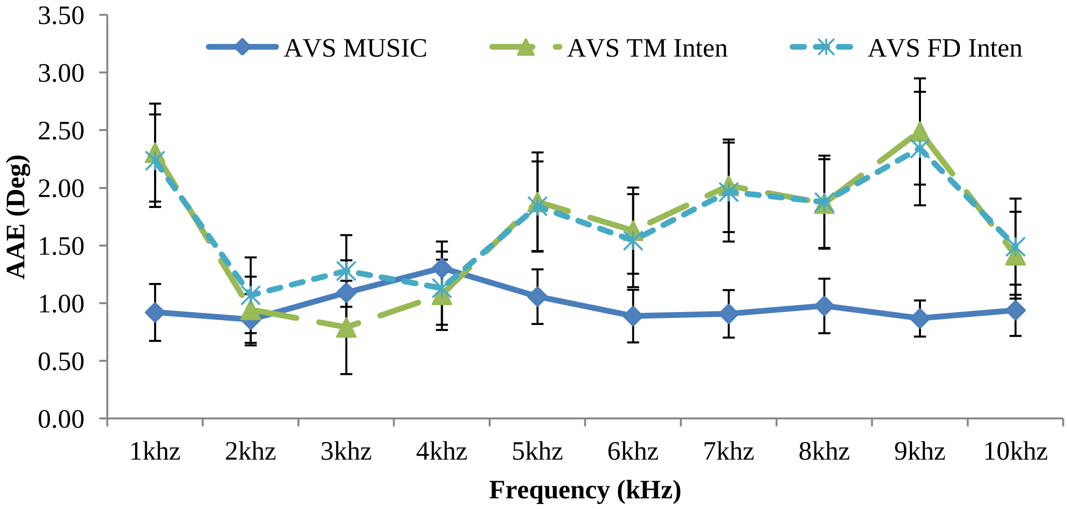


Figure 45: AAE for DOA estimates for AVS for different DOA estimation algorithms (Error bars indicate 95% confidence intervals).

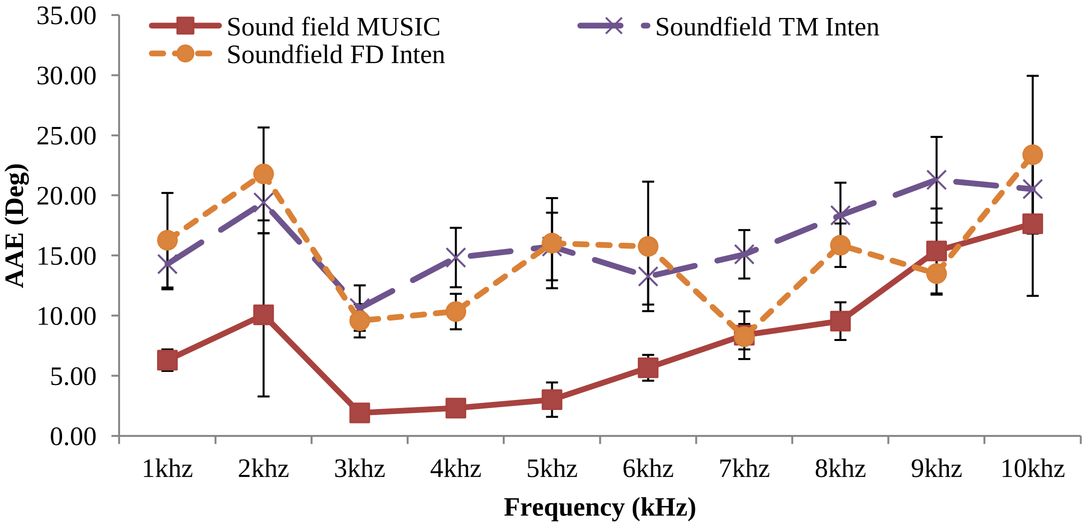


Figure 46: AAE for DOA estimates for Soundfield Microphone for different DOA estimation algorithms (Error bars indicate 95% confidence intervals).

such that the errors due to differences in gain are compensated, but due to the effects of the background noise and reflections from surrounding walls, errors are introduced. In these experiments it was found that when the noise is diffuse the error is less, but when there is a source from a particular direction (e.g. when the door of the office is open, or the phone rings) the error is higher.

The results for the Soundfield microphone are shown in Figure 46, from the results it can be seen that the average error for the MUSIC algorithm is 8.2 degree when compared to the average error for the time domain version of the intensity base algorithm which is 16.34 degrees and 15.07 degrees for the time and frequency domain implementations, respectively.

This is a doubling of the error when compared with the error from the MUSIC algorithm. Here too the intensity based algorithms rely on the gains of the pressure and directional components to vary correctly, which as explained before does not happen in reverberant and noisy conditions. The effect of the noise and reflections due to reverberations are more than that compared to the AVS.

In the case of the Soundfield array which is constructed using cardioid capsules, the outputs from the four microphones are combined according to (39) to (41). Hence, the amount of reflections and noise captured from all the directions are higher. These reflections are then included as errors in the formation of the W, X, Y and Z components. The other important factor that affects the results is the influence of the protective netting of the Soundfield as these would diffract and reflect the sound signals. In Chapter 3 it was found that for the AVS the mount and the positioning of the microphone capsules contributed to errors in DOA estimates.

In addition, for an omni-directional microphone which has no directional bearing on the output, there is relationship between the aperture of the capsule and the frequency of the signals that is if the wavelength of the signal is smaller than the aperture then the omni-directional microphone will start to display directional characteristics [13]. As seen from the results the Soundfield produces larger errors at higher frequencies especially above 8 kHz which is the frequency at which most omni-directional capsules start to exhibit the directional characteristics [13]. This change in the polar pattern may also contribute to the increase in inaccuracy of the DOA estimate from the Soundfield microphone.

4.3.3 Effect of Frame Length on the DOA Estimate

The results presented in the previous section are for average estimate of DOA of all frames, with a frame length of 20ms for a monotone signal; here the accuracy of the DOA estimate for a single frame will be investigated. Results in Figure 47 and Figure 48 are for average angular error for all DOAs for frequencies 1kHz to 10kHz for varied frame lengths from 480 samples (10ms) to 48000 samples (1s). This is done to find out

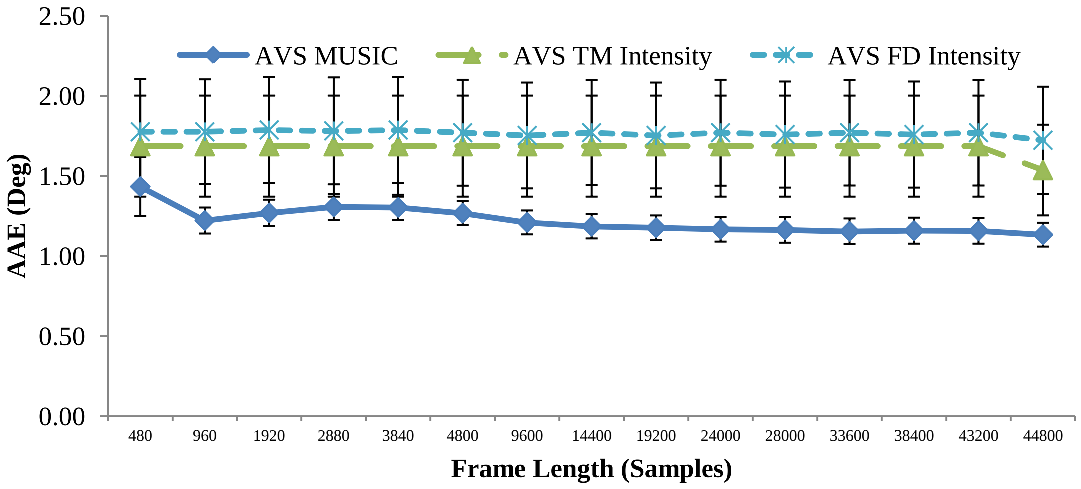


Figure 47: AAE for DOA estimates for different frame sizes for AVS for monotone signals (Error bars indicate 95% confidence intervals).

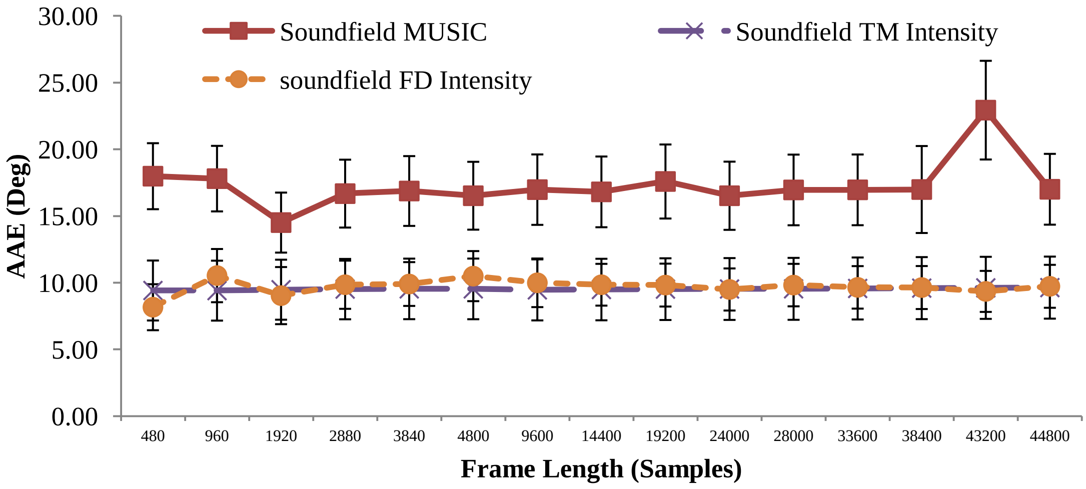


Figure 48: AAE for DOA estimates for different frame sizes for Soundfield for monotone signals (Error bars indicate 95% confidence intervals).

if it is possible to estimate the DOA from a single frame and if so what is the smallest frame length that will give accurate results.

It can be seen from Figure 47 and Figure 48, for both AVS and soundfield microphones, and for all the algorithms, the DOA estimates remains approximately equal for all frame lengths. The monotone signals have equal energy for the entire duration. Hence, the DOA estimates from single frame should be approximately the same as that of the average.

In real applications signals such as speech may have a time varying energy, and so, the DOA estimates from each frame may be different, especially if the source is moving. Hence, it is crucial to find the smallest frame length at which an effective DOA estimate can be obtained for a single frame of speech.

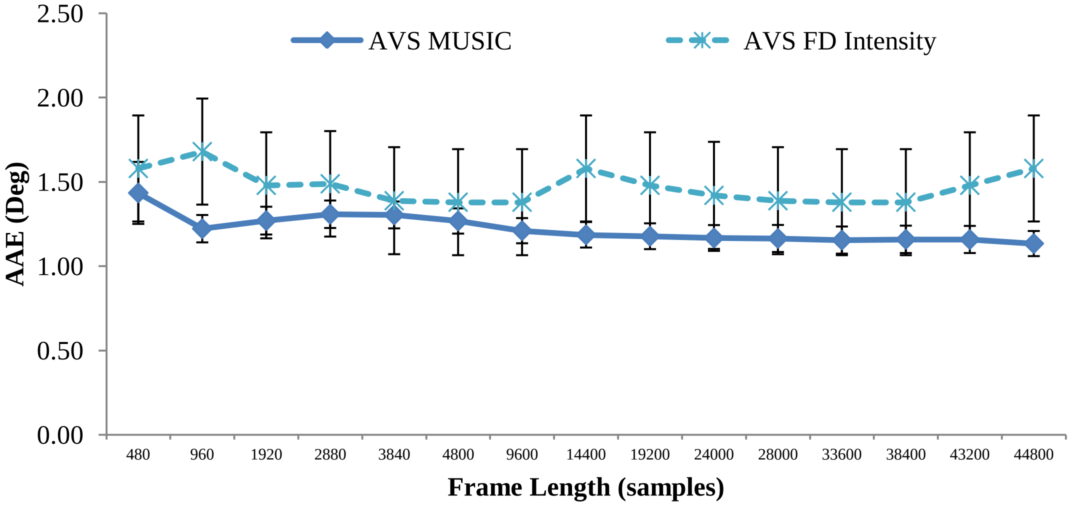


Figure 49 : AAE for DOA estimates for different frame sizes for AVS (speech signals) (Error bars indicate 95% confidence intervals).

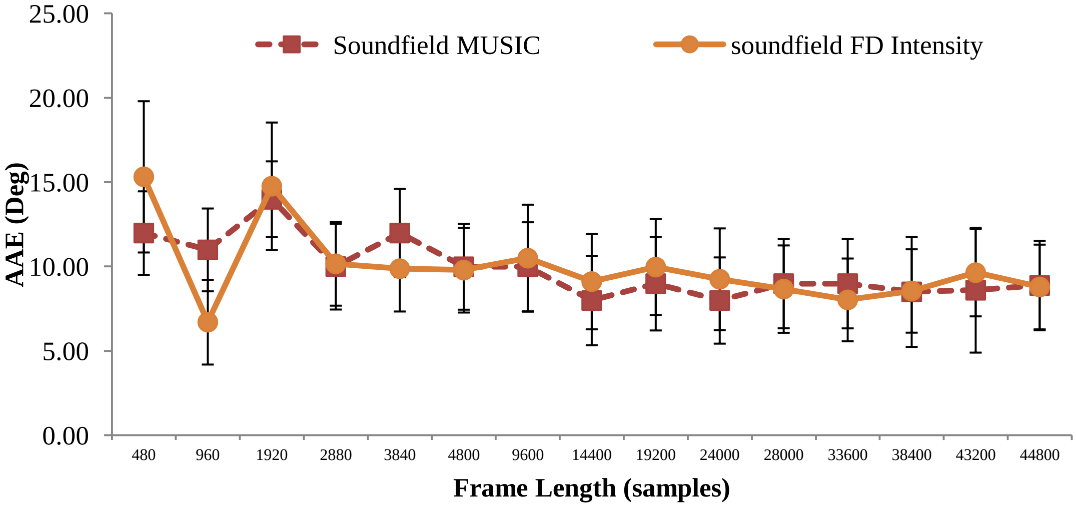


Figure 50: AAE for DOA estimates for different frame sizes for AVS (speech signals) (Error bars indicate 95% confidence intervals).

The results presented in Figure 49 and Figure 50 are for different frame lengths from 10ms to 1s for speech signals; here, the speech test signals are deliberately created such that the entire speech frame is voiced (that is all the unvoiced section are artificially removed). The results show by changing the frame length there is no change

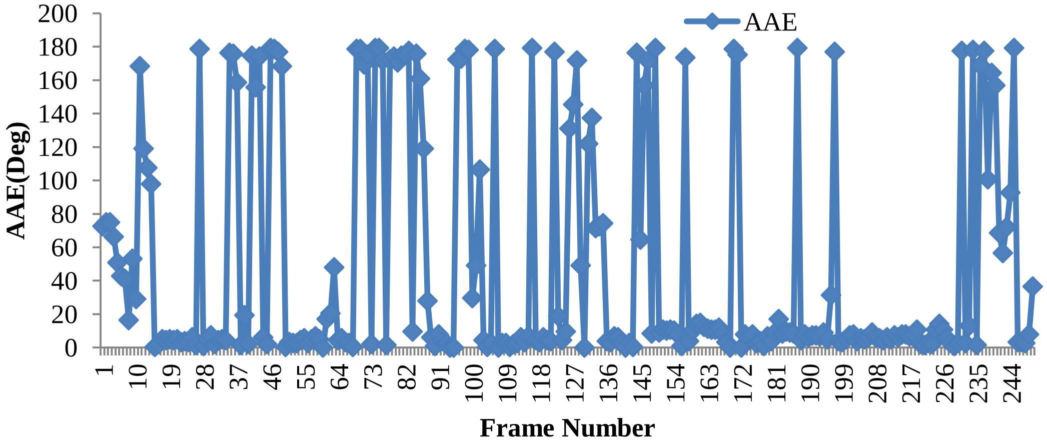


Figure 51: AAE for DOA estimate for each frame of a speech sentence.

in the performance for any of the algorithms, this result shows that it is possible to get an accurate DOA estimate for frame size as small as 10ms which is the frame length used in most real time speech applications. In these results the time domain implementation of the intensity based algorithm is not included as it was found that this algorithm failed to give any statistically consistent result for speech sources.

4.4 Stationary Speech Sources

Unlike monotone signals, speech signals have different characteristics. The energy of the speech signal varies over time, there are voiced, unvoiced and silence in the sentence which should be considered. From the previous section it has been established that frame lengths of 10ms is enough to get an accurate DOA estimate for a speech signal. The results presented in Figure 51 are for all the frames of a speech sentence with the speech source located at 0 degrees in azimuth to the microphone with a frame length of 480 samples or 10ms using the speech test database described in Section 4.3.1. The results show that all the regions of the speech which are unvoiced or stops produce errors and the AAE is 49 degrees.

This is expected as these regions are affected more by the noise form background. These results show that in practice for DOA estimation of speech like signals, which contain voiced, unvoiced, and stops, an accurate DOA estimate cannot be obtained by averaging the DOA estimates from all frames. Furthermore it is important to distinguish between speech, which are voiced and those that are unvoiced and stops before calculating the DOA for that particular frame.

4.4.1 Voice Activity Detection

To identify if a frame is voiced or unvoiced, a Voice Activity Detector (VAD) can be used. This is an important feature in most telecommunications systems where it is important to identify if a frame is voice or unvoiced in terms of reducing the bit rate, saving power of mobile devices, reducing co-channel interference in mobile devices and greater noise suppression in speech enhancement [140].

The basic idea behind a VAD is to analyse the expected value of Power Spectral Density (PSD) of overlapped frames. The comparison is made between the PSD of a noise frame and frame with noise and speech. A statistical likelihood ratio between the PSD of noise only frame and the frame with noise and speech is made and statistical bayes test is carried out by comparing likelihood ratio against a predetermined threshold. The basic idea of most VAD is the same, but the way in which different techniques calculate the thresholds determines the accuracy of the VAD [140].

In this thesis, the VAD based on ITU-T G.729B [141] is used. The frame length used here is 10ms, which is the smallest frame length tested in the previous section, where an accurate DOA estimate is obtained. Furthermore, when the frame length is 10ms it is assumed that the voiced and unvoiced sections of the speech can be identified efficiently and there is no significant change in the energy of the speech in that frame.

4.4.2 DOA Estimation with VAD Incorporated in the DOA Algorithm

Results in Figure 52 and Figure 53 are for the DOA estimation for speech sources with a VAD implemented in the algorithms. The results show that with the VAD in place the AAE for the AVS is 1.58 degrees from the MUSIC algorithm and average error for frequency domain intensity algorithms is 1.57 degrees. The results in Figure 52 show that the errors between the different algorithms are small and furthermore the error bars overlap for all angles. Hence it can be concluded from this result that the difference in performance for different algorithms when applied to real speech recordings is negligible and the three algorithms perform equally.

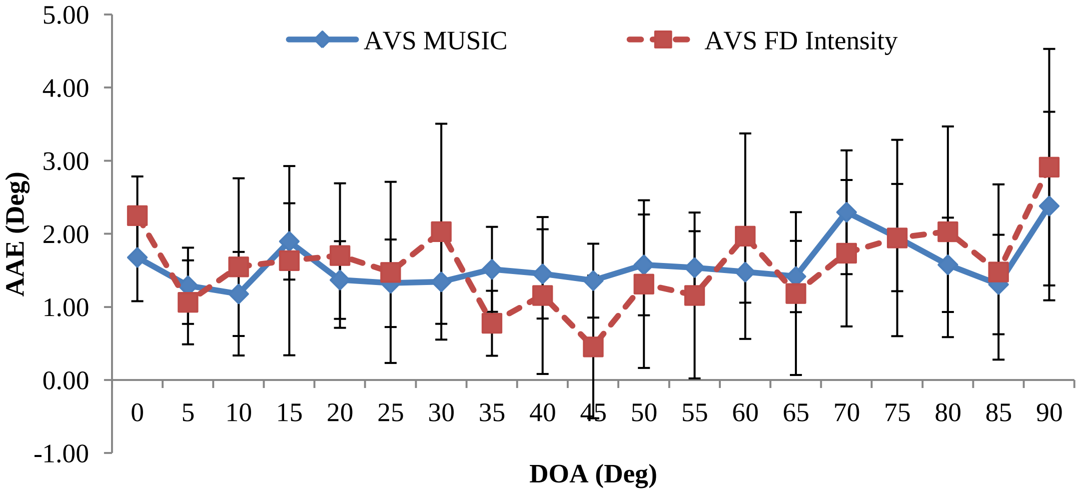


Figure 52 : AAE for DOA estimates for different frame sizes for AVS (Error bars indicate 95% confidence intervals).

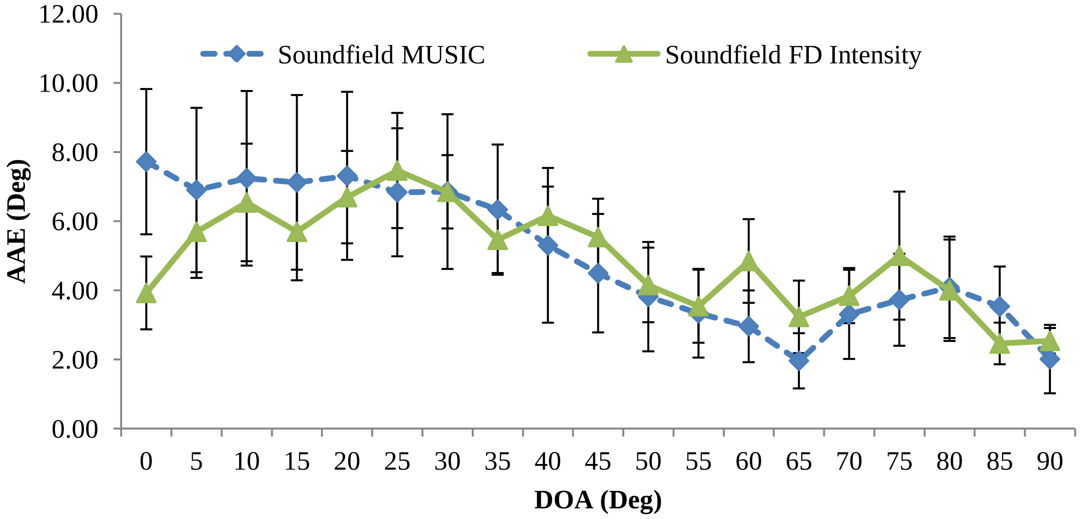


Figure 53: AAE for DOA estimates for different frame sizes for Soundfield (Error bars indicate 95% confidence intervals).

The results presented in Figure 53 are for the DOA estimates of speech from Soundfield microphone. The results show that the average error for the MUSIC algorithm is 4.99 degrees, and the error for the error for frequency domain intensity based algorithm is 4.93 degrees. The results show that for the Soundfield, the MUSIC algorithm and the frequency domain version of the intensity algorithm are approximately equal. Since most of the error bars overlap it can be concluded that statistically the results are equal for the three algorithms.

The effect of using the VAD to filter out frames that that are unvoiced is clear from this result, as can be seen when frames without voice are used in the calculation of the DOA results have a higher error than when only voiced frames are used.

4.4.3 Moving Speech Sources

The results for stationary sources were presented in the previous section. In this section DOA estimates for moving sources will be presented. The importance of the ability to estimate DOA estimates for moving sources is for applications such as automatic camera panning in video teleconferencing, where when a client on one end moves during a presentation the camera is able to follow the moving speaker.

	Slow		**Normal**		**Fast**	
	MUSIC	Intensity	MUSIC	Intensity	MUSIC	Intensity
0	2.36	3.71	5.39	4.78	1.05	6.09
30	3.96	3.30	3.98	5.13	4.94	1.47
60	5.72	5.69	5.50	1.11	3.41	0.46
90	2.91	4.22	2.50	4.41	1.58	5.00

Table 3: AAE of MUSIC and Intensity algorithm for moving source for AVS

The results presented in this section are for the three algorithms and for a source moving at three different speeds, which are slow, normal and fast moving speakers. The time taken for an average person to walk an arc of 10 degrees at a distance of 1m from the microphone is 0.13sec, which is 13 frames at 48 kHz sampling rate and frame sizes of 480 samples. The time taken for a person moving through a 10 degree arc is larger than the frame length required for producing an accurate DOA estimate. But because the speech has unvoiced sections and stops, a more reliable estimate can be obtained by using as many frames as possible. Hence, the length of the speech segments in each speaker is at least 6 frames long and the time taken for the speech segment to move from one loudspeaker to the next is described below.

To simulate moving targets, three additional loudspeakers were used as shown in Figure 44. The average speed of walking for a human being is 1.33m/s. This means on average in a circular path with a radius of 1m a person walking at this average speed would take 0.13s to walk 10 degrees. The speech sentences were sliced into four parts each part 0.066s long for fast moving, 0.13 s for normal walking speed and 0.3 s for slow walking paces and the speakers are separated by 30 degrees.

	Slow		Normal		Fast	
	MUSIC	Intensity	MUSIC	Intensity	MUSIC	Intensity
0	8.21	16.10	43.11	15.28	6.95	21.02
30	66.32	17.23	110.29	19.27	119.41	13.02
60	41.81	21.47	45.00	14.59	74.01	16.60
90	5.03	8.09	4.56	25.23	13.40	7.70

Table 4: AAE of MUSIC and Intensity algorithm for moving source for Soundfield

Each part of the sentence is played on one loudspeaker in order and between each part a silence of approximately 0.2s for fast moving, 0.4s for average walking speed and 0.8s for slow walking is introduced. Hence, the experimental setup simulates a source moving over 4 sectors, each covering 10 degrees.

The results presented in Figure 54 to Figure 56 are for those of a source moving at slow, normal and fast walking speed recorded by an AVS. The results in Figure 54 to Figure 56 show that all algorithms give accurate DOA estimates for all three walking speeds. The AAEs for the results presented in Figure 54 to Figure 56 are given in Table 3. The results of Table 3 show that the results obtained from the MUSIC algorithm has less error than the results from the intensity based algorithms. In all the experiments performed, the results have shown consistently that the errors from the MUSIC algorithm are smaller than that of the intensity based algorithm.

The results presented in Figure 57 to Figure 59 are for the recordings of the Soundfield microphone for sources moving at the three speeds. Unlike the AVS the errors from the Soundfield microphone are seen to be higher for all the speeds and especially the DOA estimates from the MUSIC algorithm is higher for the Soundfield compared to the intensity based algorithms. The AAE for the three speeds for the Soundfield microphone are given in Table 4.

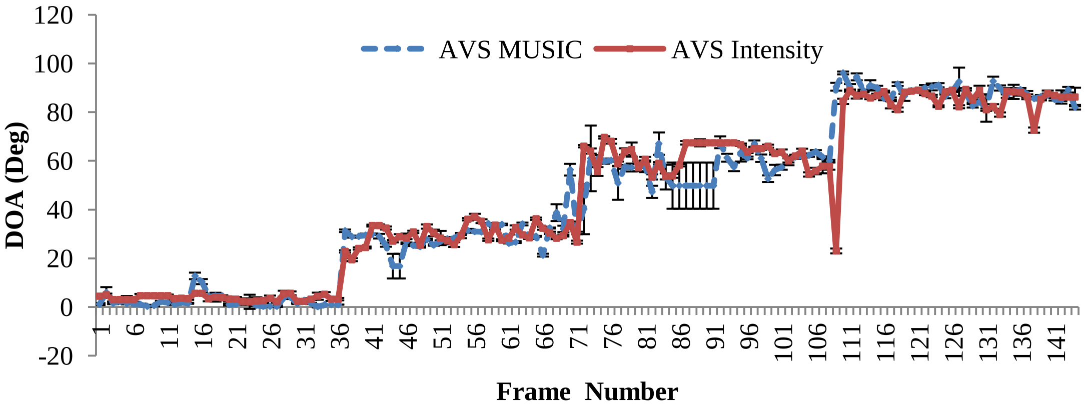


Figure 54: DOA estimate for slow moving speech source form AVS (Error bars indicate 95% confidence intervals).

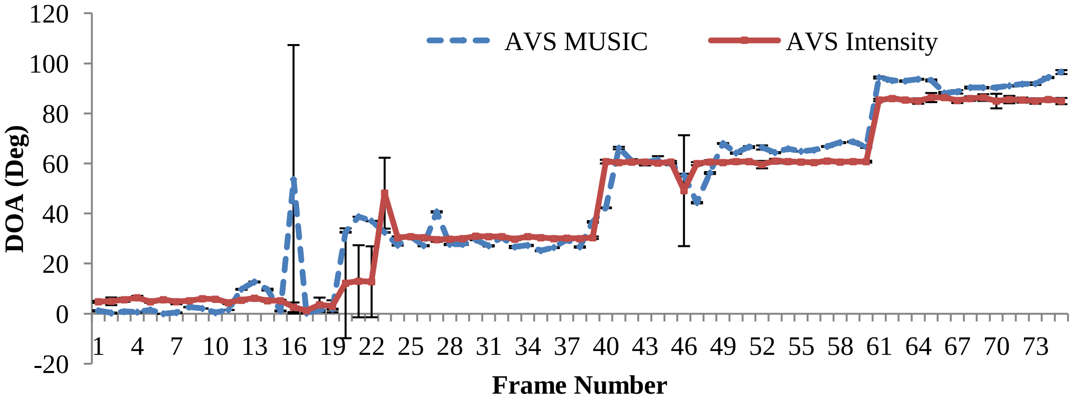


Figure 55: DOA estimate for normal moving speech source form AVS (Error bars indicate 95% confidence intervals).

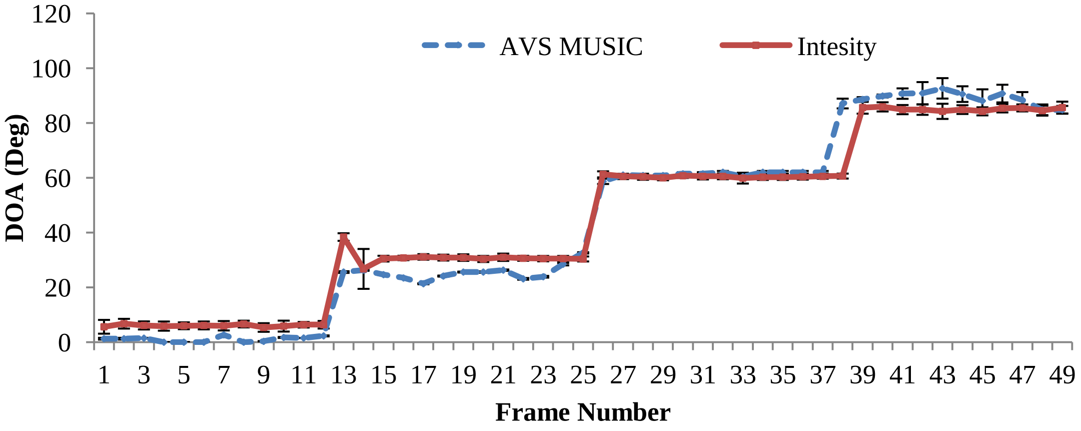


Figure 56: DOA estimate for fast moving speech source form AVS (Error bars indicate 95% confidence intervals).

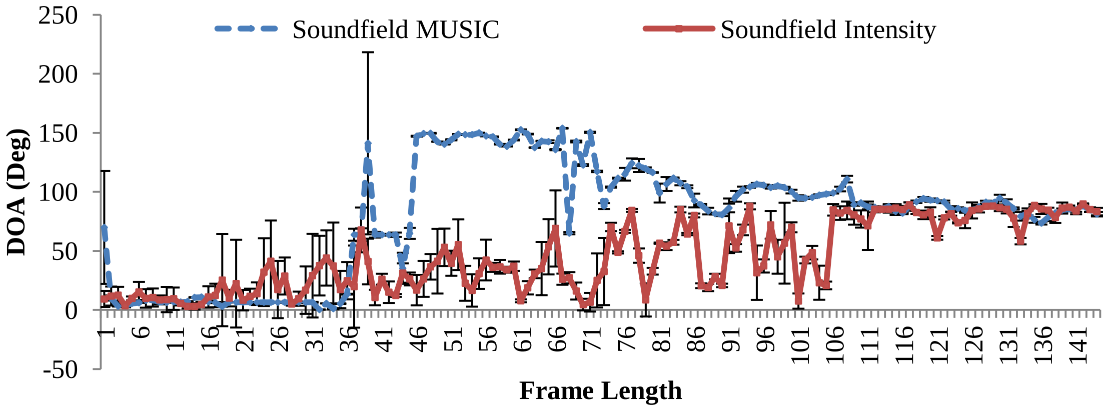


Figure 57: DOA estimate for slow moving speech source form Soundfield (Error bars indicate 95% confidence intervals).

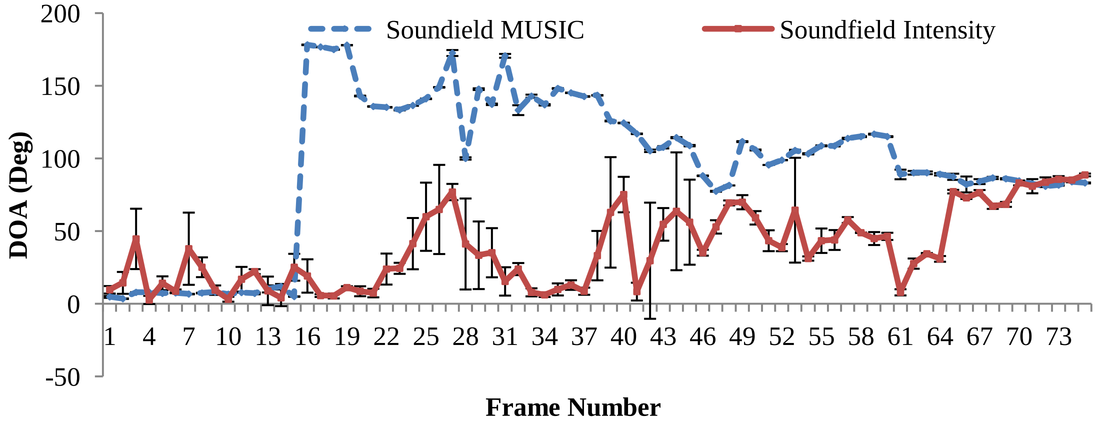


Figure 58: DOA estimate for normal moving speech source form Soundfield (Error bars indicate 95% confidence intervals).

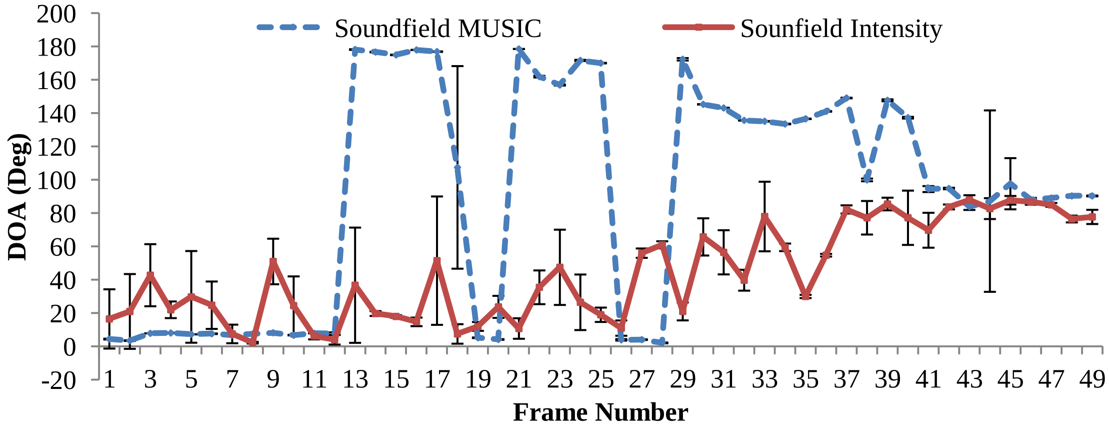


Figure 59: DOA estimate for fast moving speech source form Soundfield (Error bars indicate 95% confidence intervals).

4.5 DOA Estimation for Multiple Sources

The algorithms presented so far have shown good performance for a single source case. When the number of sources increases the task of determining the DOA estimates becomes harder and much more challenging. In the current literature, there are many different approaches for obtaining DOAs from multiple sources, which include the use of a BSS algorithm first to separate the mixed sources into individual components and then to obtain the DOAs for different components as proposed in [142-145], and the use of clustering techniques with existing DOA estimation methods as in [46, 48, 146-150]. The approach of BSS for DOA estimation have several problems, which include complexity of BSS algorithm and the amount of data needed for BSS to work efficiently. In theory, the MUSIC algorithm in its basic form is able to provide DOA estimates for multiple sources. The time domain intensity based algorithm when used to obtain a DOA estimate for speech sources failed to give valid results. Hence, in the case of the multiple sources this algorithm will not be used.

Unlike the time domain version of the intensity based algorithms the frequency domain intensity based algorithm calculates the DOA for individual frequency bands. Hence, if the frequency content of the sources is in different frequency bands then in theory the DOA estimate for each of those bands should correspond to an individual source. This idea relies on the sparsity of speech in the time-frequency domain, where multiple simultaneous speech sources have minimal overlap in this domain. Hence, each time-frequency component will in general belong to one speech source. In order to calculate the DOAs of sources which have frequency components that are close together the width of the FFT bins must be smaller.

In addition to the frequency components of the sources, the frame length and individual frames used in the processing plays an important role in the accuracy of the DOA estimates. Unlike the case when there is a single source the DOA estimates from different sources will produce different DOAs for different frequency bins and for different frames, hence, it is important to analyse each frame to indentify how many sources are present and which DOAs are due to errors. Furthermore, the frame length must be as small as possible such that changes in the DOAs can be obtained accurately. Hence, the DOA estimates from the frequency domain algorithms can be analysed in two parts which are:

- DOAs for each frequency bin in each frame
- DOAs from multiple frame

For a real time application the DOAs from different frequency bins of a single frame is more important than combining DOAs from multiple frames, but for increased accuracy, combining multiple frames will give better results. Here, both approaches will be analysed.

Since, the number of frames and number of FFT coefficients are extremely large, there are number of methods that can be used to analyze the data of DOA estimates from the FFT coefficients and form the frames. One of which is data clustering as described above. Hence, a brief discussion on different methods clustering is presented next.

4.5.1 Data Clustering

The definition of clustering according to [151] is unsupervised grouping of similar objects, which means that different clusters will contain objects that are different. The similarity between two data points can be measured by a distance measure, which measures how close the two data points are, this type of clustering is known as distance based clustering [151], or by grouping the data points based on the data types. In general, clustering of data can be performed in two broad methods which are partitional and hierarchical clustering algorithms.

4.5.2 Hierarchical Clustering

The [151] hierarchical clustering is based on recursively assigning the data points into clusters. The general form of hierarchical clustering can be explained according to the following steps as described in [151, 152]:

1) Each data point is assigned to a cluster (e.g.: if there are N data points, N clusters are formed with just one data point) and the distances between the clusters is assumed to be the same as the distances between the data points.
2) The closest pair of clusters are found and merged into a single cluster, hence, now there are $N-1$ clusters.
3) The distances between the new cluster and the old clusters are updated.

4) Steps two and three are repeated until distance between the clusters is more than a set threshold.

There are three forms of hierarchical clustering, which are [152]

1) Single-link – the distance between the clusters are assumed to be the shortest distance between any member of one cluster to any member of the other cluster
2) Complete-link – the distance between the clusters are assumed to be the longest distance between any member of one cluster to any member of the other cluster
3) Minimum-variance - the distance between the clusters are assumed to be the average distance between any member of one cluster to any member of the other cluster

The advantage of using a hierarchical structure is that there is no requirement of how many clusters should be formed and the drawback of the hierarchical algorithms is when there is a large data set these algorithms suffers due to the recursive nature of the algorithm.

4.5.3 Partitional Clustering

The partitional clustering algorithm forms a set number of clusters, and then assigns the data points into those clusters. One of the problems of having to determine a set number of clusters is that in real applications such as DOA estimation the number of clusters is unknown [153]. The partitional techniques usually produce clusters by minimizing a criterion function defined either locally using, e.g., as Probability Density Function (PDF) functions or globally, such as minimizing the distance function within the clusters and maximizing the distance function between clusters. Due to the large number of combinations that are possible for assigning the data point to a cluster, the algorithms run multiple times to get the best possible configuration of the clusters. One of the most often used partitional clustering algorithm is the *k*-means algorithm, where the criterion used is the average squared distance to the centre of the nearest cluster [153]. The *k*-means algorithm starts with randomly assigned cluster centers and assigns the data points to the closes centers, then the centers are updated and data points are reassigned. This process is repeated until convergence is achieved. The convergence

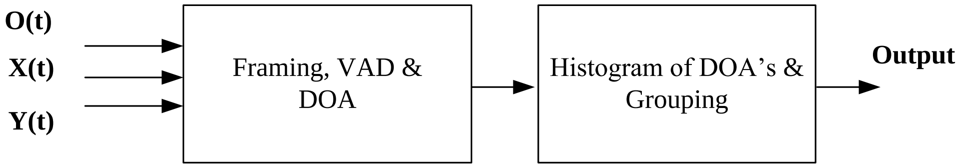


Figure 60: Block diagram of the proposed method.

occurs when data points are no longer assigned or there is minimal decrease is squared error [153].

4.5.4 The Format of DOA Data

The implementation details of the two algorithms presented for a single source has been discussed before. For a single source, the frequency domain version of the intensity algorithm produces N number of DOA estimates where N is the number of FFT bins, whereas the MUSIC algorithm outputs multiple DOAs for each frame. When the number of sources is more than one for each frame, the MUSIC algorithm and the frequency domain intensity algorithm may produce more than one DOA estimate per frame. In contrast for a single source, a simple averaging of the DOAs from each frame gives an accurate DOA estimate as explained in Section 4.3.

To analyse the DOAs for multiple sources, a simple averaging of the DOAs from each frame or by averaging the DOAs from all the frames will produce errors. Hence, a different technique is required. The most complex data structure is produced by the frequency domain version of the intensity algorithm, also the analysis technique for this method is presented first.

4.5.5 A Method for Analysing the Output from Frequency Domain Intensity Algorithm

The DOA estimation approach of Section 4.2 results in a direction estimate for each time-frequency component. This section describes the method for estimating DOAs for mixed speech sources by combining time-frequency components with similar direction estimates. A block diagram of the process is shown in Figure 60. The approach used here is based on the clustering techniques described in Sections 4.5.1 to 4.5.3. From the discussion in Sections 4.5.1, it is clear that the best approach for analysing the DOAs from each frame is by using a clustering technique. But since the number of FFT point for each frame is at least 512, the hierarchical structure will not be the best choice due to the recursive nature of the algorithm. The partitional methods, on the other hand, require that the number of clusters or number of sources be known which in real applications is not the case. Hence, a clustering technique which does not require the knowledge of the number of clusters, and that does not require a recursive sorting technique as that of the hierarchical structure is required. Such a technique is presented next.

The speech signals from the AVS are formed into 10ms frames with an overlap of 50% using a Hamming window. After framing, the frames are passed through a VAD. The VAD used in this work is based on a modified version of the VAD of ITU-T G.729B [141]. If the frame contains active speech, then the frame is passed to the DOA algorithm, the FFT of the frame is taken, and the DOA estimate for each frequency bin found using the intensity approach of (130).

Let the space around the AVS in azimuth from 0 to π degrees be divided into 5 degree intervals (this is the resolution used in the DOA estimation in Chapter 3 and for single source in Section 4.3 and 4.4 of this chapter). Now a matrix **U** of size 2×*36* can be formed as shown in (131).

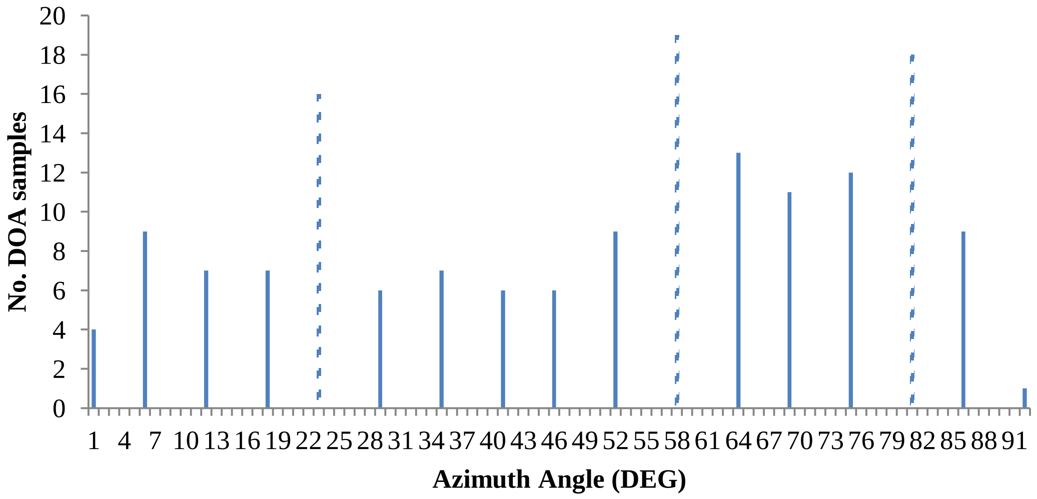
$$\mathbf{U} = \begin{bmatrix} his_1 & his_2 \; ... & his_{36} \\ mid_1 & mid_2 \; \cdots & mid_{36} \end{bmatrix} \tag{131}$$

$$id_m = \left[k_1 \cdots k_{his_n}\right] \tag{132}$$

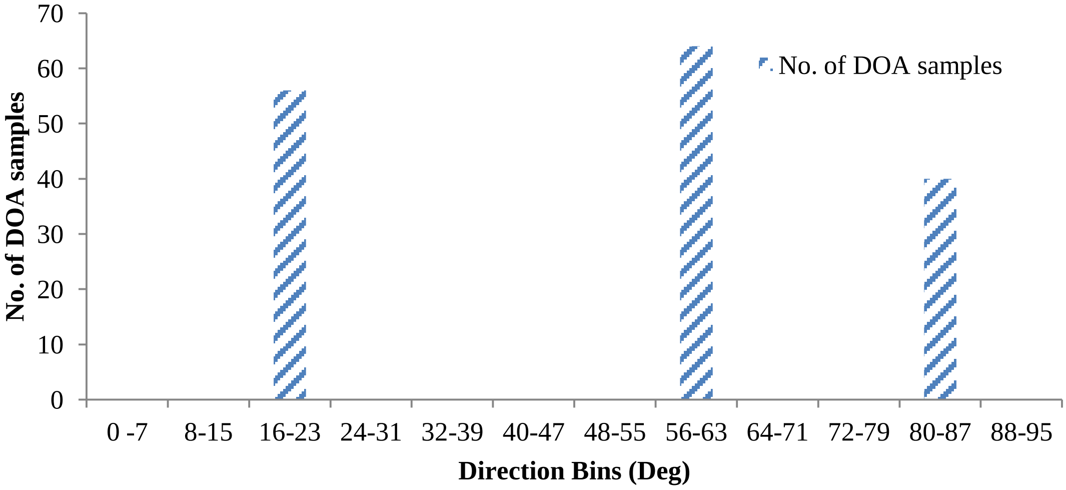
where his_m is known as a direction bin and contains the count of DOA estimates from (130) that fall into the m^{th} interval (there are 36 intervals between 0 to π), and where id_m is a vector of the indices representing the set of frequency components that produced the DOA estimates that fall into the bin corresponding to his_m and mid_m is the midpoint of each m^{th} bin interval. Figure 61 (a) shows the plot of the first row of **U** for an example time-frequency frame from a recording of three simultaneously occurring speech sources.

The elements of the first row of **U** are sorted and the largest peak is identified as the first source. The remaining unique sources are identified by comparing the remaining histogram peaks with the largest peak. A new unique source is found if the expression of (133) is true.

$$\frac{\boldsymbol{U}(1,m)}{\boldsymbol{max}(\boldsymbol{U}(1,m))} \geq \gamma \qquad m = 1, \dots, 36 \tag{133}$$



(a)



(b)

Figure 61: Histogram of time-frequency direction estimates for a frame derived for an example recording of 3 mixed sources (a) Original histogram, with peaks of each source indicated by lighter shading (b) Histogram following sorting and clustering of direction estimates corresponding to each source.

In the current work, it was found that γ = 0.60 produced the best results; the experimental evaluation of the best value for γ is presented in the next section. In practice, this parameter could also be interactively adjusted by a user to provide increased or decreased accuracy of DOA estimates of desired speech signals. As illustrated in Figure 61 (a), three peaks are identified for the three sources of this example mixed speech frame. The remaining direction bins that are below the threshold

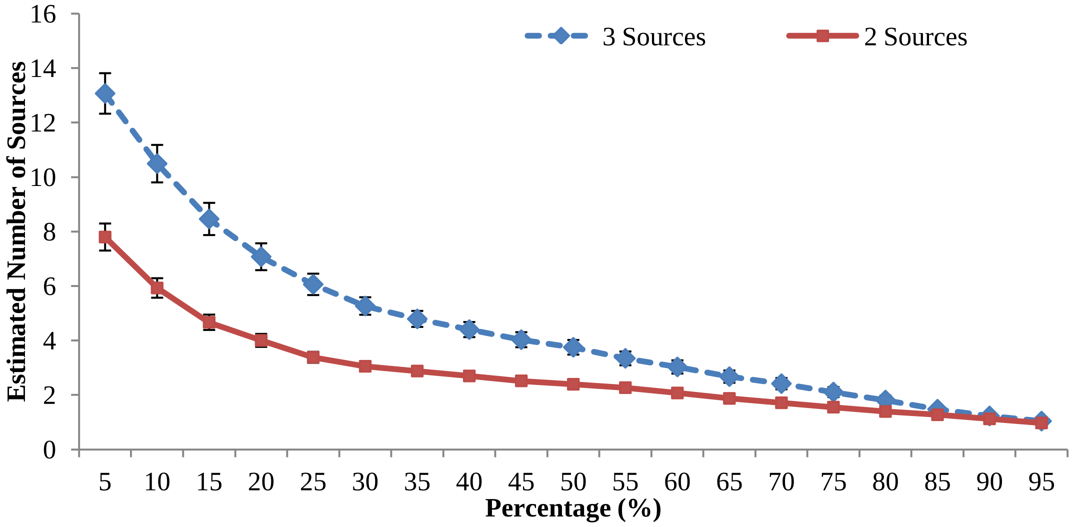


Figure 62: The graph of threshold values against the estimated number of sources for two and three sources (Error bars indicate 95% confidence intervals).

of (133) are deemed to be due to errors in DOA estimation, secondary reflections or time-frequency components belonging to more than one source. For these remaining histogram bins, a clustering approach is applied, whereby the direction of the source for these bins is assigned as the direction of the closest histogram peak. For the remaining direction bins, if there are three sources which are $\mathbf{U}_{(1,3)}$,$\mathbf{U}_{(1,8)}$ and $\mathbf{U}_{(1,11)}$ as illustrated in Figure 61. Then for each source, the distance between the remaining direction bins are found:

$$\min\,(\text{distance}) = mid_{(s)} - mid_{(m)} \tag{134}$$

where $m \neq (3{,}8\ and\ 11)$ and $s = 3{,}8$ and 11, for the direction bins that produces the minimum distances , the contents of the $id_{(m)}$ that satisfy (133) is copied to $id_{(s)}$.

4.5.6 The Experimental Evaluation of the Threshold Value for Two and Three sources

The methods for analysing the DOAs from a single frame described in the previous sections require a threshold to identify the possible number of sources. The threshold is the value of γ in (133) which is used to indicate a unique source. Recordings were made of two and three consecutive speakers according to the setup of Figure 44 and according to the description of Section 4.3, in total 135 recordings were made for three sources and 180 recording were made for two sources.

A 2048 point FFT is performed on each frame, which gives a resolution of 4 Hz at a 16 kHz sampling rate. The results presented in Figure 62 show the number of sources chosen for different values of the threshold, for all the sample files tested. The results in the Figure 62 shows when the threshold is high, the number of sources are reduced and when the threshold is low the number of sources increase. Since these results are obtained from 315 recordings, it can be safely assumed that the threshold value obtained from the experimental procedure is valid. At $\gamma = 60\%$ the algorithm correctly predicts the number of unique sources for two and three sources.

4.5.7 Results for DOA for Multiple Sources from Time Domain MUSIC Algorithm

The recording of two and three sources was processed using the MUSIC algorithm. The results obtained from these recording showed that for each frame a single DOA estimate is produced. This DOA represented different sources in different frames and huge variation in the errors for the DOAs were obtained. Hence, for a single frame the errors were found to be statistically invalid.

To evaluate the results further; 200 fames are grouped and the clustering technique described in Section 4.5.5 was applied. The results obtained from this process are shown in Figure 63 and Figure 64. The results show that for the two sources the DOA estimates from the clustering technique of Section 4.5.5 does identify two sources, but the errors obtained are very large. For the case of three sources, the proposed clustering technique only identifies two unique sources, and like the case of the two sources the errors are very large hence the results are not statistically reliable.

4.5.8 Results for DOA for Multiple Sources from the Frequency Domain Algorithm

The outputs from the AVS for two and three consecutive speakers were processed using the techniques described in Section 4.5.5. The first experiment conducted in this section is to identify the effect of FFT length on the accuracy of the DOA estimation. In Chapter 2, the human speech production mechanism was discussed where it was identified that what separates two individual speakers is the resonant frequency which is the F0 (the first harmonic). The ranges of these F0 for male and

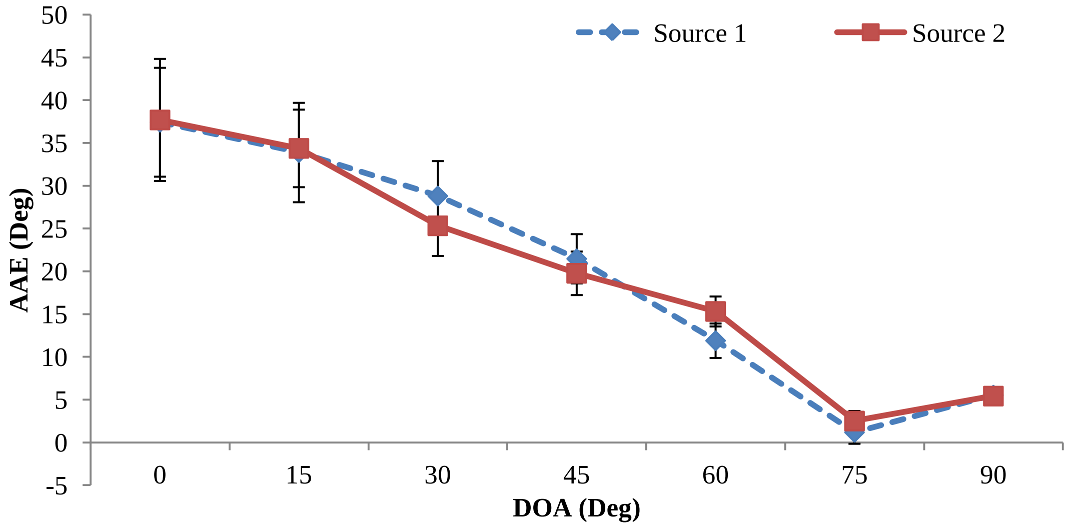


Figure 63: AAE for DOA estimates from MUSIC algorithm for two sources. (Error bars indicate 95% confidence intervals)

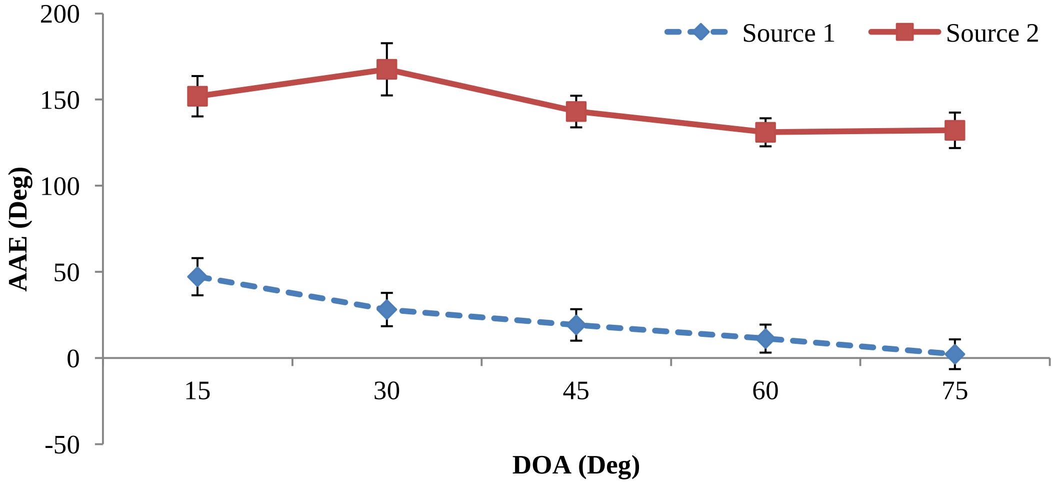


Figure 64: AAE for DOA estimates from MUSIC algorithm for three sources. (Error bars indicate 95% confidence intervals).

female speakers were identified in Section 2.9.3 to be between 60 and 400Hz. Hence, to identify two individual consecutive speakers the width of the FFT bin must be narrow enough to distinguish between two F0. The recordings used in this work are down sampled to 16 kHz, and if the FFT length is set at 512, then the resolution of FFT bins is 31.5Hz, which means if there are two speakers with F0 of 120Hz and 130Hz, then only one DOA estimate will be calculated for both speakers.

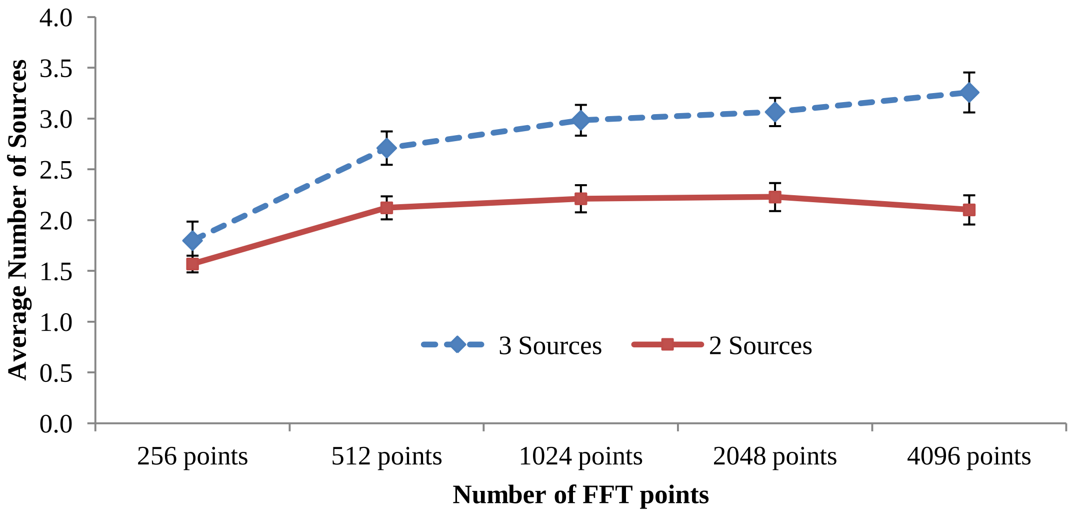
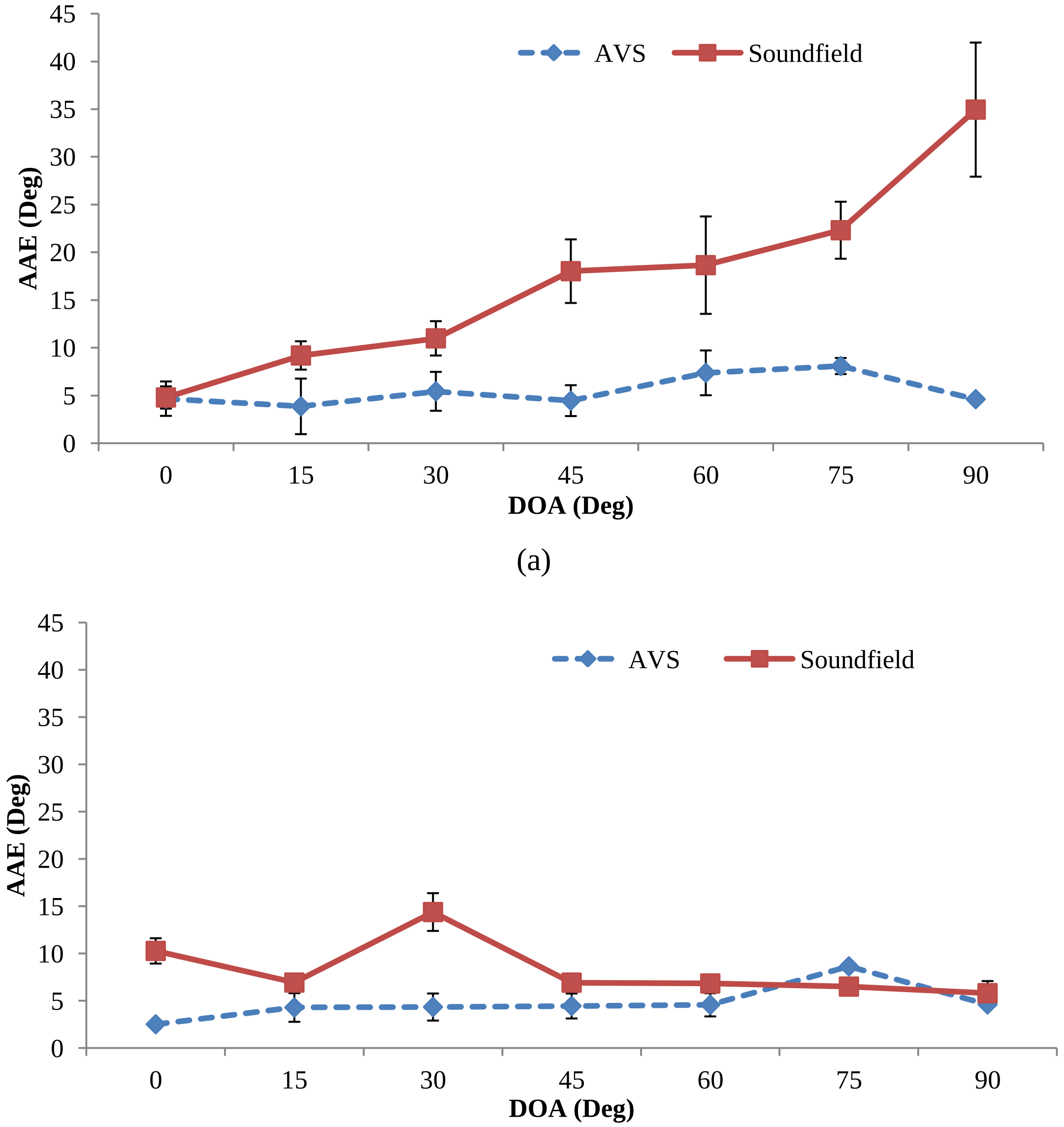


Figure 65: The relationship between the number of sources obtained from and the number of FFT points. (Error bars indicate 95% confidence intervals).

The results presented in Figure 65 show the relation between the number of FFT points and number of sources identified. The database used to generate the results consists of both male and female speakers, hence, contains all male, all female and male and female speakers speaking, consecutively. From the evaluation of the results it is seen that in general the error in identifying the number of speakers are higher, when the number of FFT points are less than 2048.

For recording where there is a mix of both male and female speech, the errors are smaller for 512 point FFT and 1024 point FFT. Overall, when the number of FFT points is less than 512 for most files, the algorithm failed to identify 3 sources correctly, and in most cases for three sources the algorithm identified only two sources or one source. From these results it can be concluded that to obtain the correct DOAs, the number of FFT points must be greater than 512, where 512 is the smallest length of FFT that will give acceptable results.

Since these results were obtained for a threshold of 0.60, by reducing the threshold it is possible to improve the accuracy when implementing the algorithm with a 512 point FFT. Based on these results the algorithm for estimating the DOAs was implemented with a 2048 point FFT. The problem with using a longer FFT length is reduced efficiency of the DOA estimation algorithm. In applications where a rough estimate (e.g.: for source separation in real time) of the DOA is only required a shorter FFT length can be used to get a faster processing time.

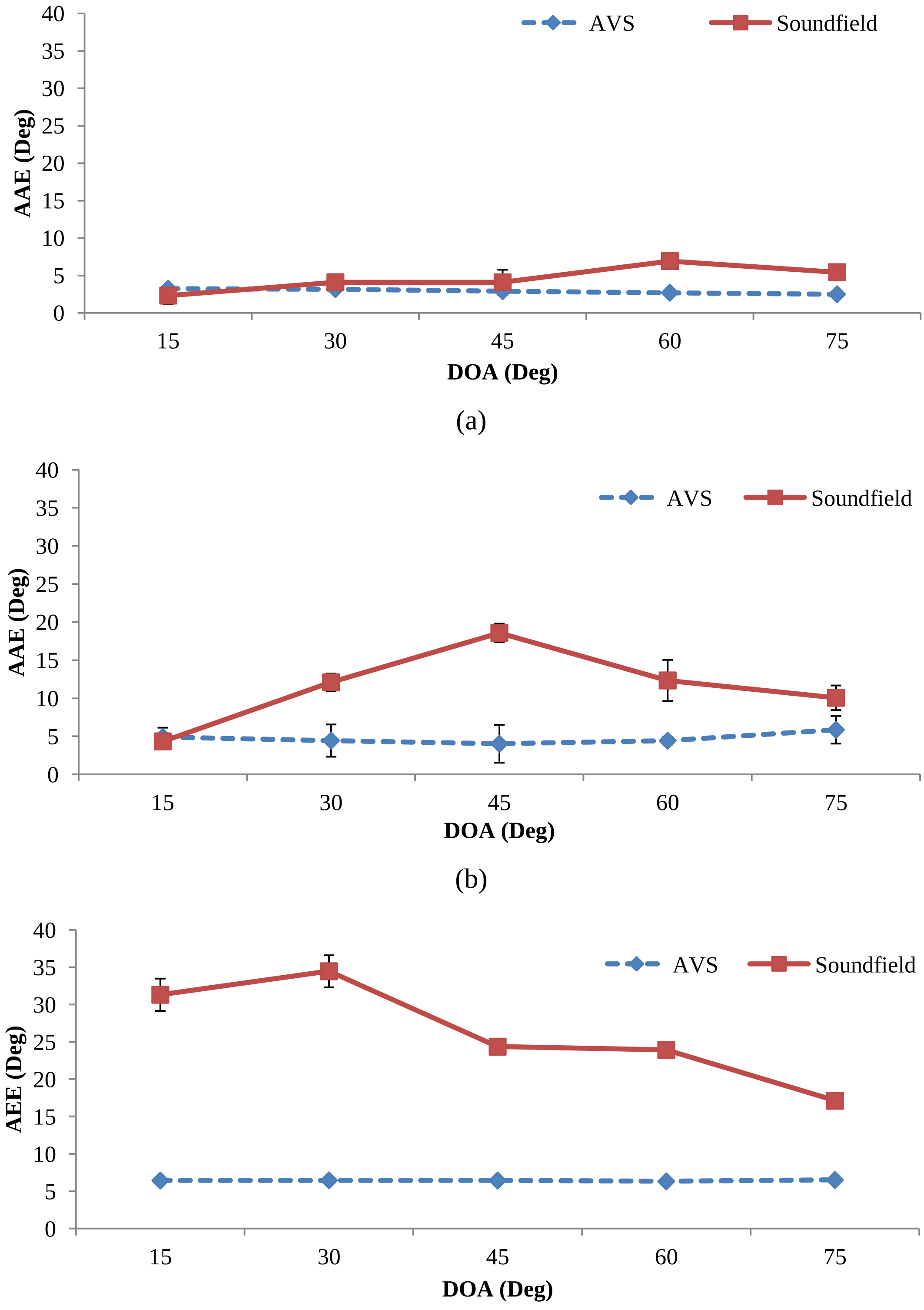


(a)

(b)

Figure 66: The AAE Vs DOA for two sources, a) Source 1 b) Source 2 (Error bars indicate 95% confidence intervals).

The results for DOA estimation for two and three sources are presented in Figure 66 and Figure 67 using FFT of 2048. Compared to the results of the single speech source, the accuracy of the DOA estimate suffers when there is more than one source. The average results for all files shows an average AAE of 5 degrees for all DOAs for two sources, and the maximum AAE can be as high as ±7 degrees, which means the actual error for any given sample could be as high as 7 degrees. In the case of three sources the results presented show that on average the AAE for the source at 90



(a)

(b)

(c)

Figure 67: The results for AAE vs. DOA for three sources a) Source 1 b) Source 2 c) Source 3 (Error bars indicate 95% confidence intervals).

degrees in azimuth is larger than the other sources. Here too the average AAE for all the sources is on average 5 degrees.

The algorithm used for the AVS is also applied to the recording from of the Soundfield microphone. The AAE for the recordings of two sources from a Soundfield microphone is 12.6 degrees and for three sources is 14.9 degrees. The errors for the Soundfield microphone are higher compared with the AVS, and are consistent with the errors that were obtained for the Soundfield for a single source, but the errors for multi source for the Soundfield microphone is much better than the single source case. This improvement in results is because of the sorting algorithm, which provides a much more accurate result in DOA estimation compared to averaging of the DOA estimates. As explained before, the higher errors for the Soundfield microphone are due to the fact that the Soundfield microphone captures more reverberation and noise compared to the AVS.

These results show that with a directional microphone array such as the AVS, it is possible to get an acceptable DOA estimate for multiple speech sources. A further test was carried out to see if the performance of the system improved when a conventional clustering technique is used. The results for applying the *k-means* clustering technique to the DOA data from (130) are presented in the next section.

4.5.9 DOA Estimation Using the *k-means* Clustering

The *k-means* clustering technique is one of the most well known clustering techniques that have been used in the field of data mining. As explained in section 4.5.3, one of the disadvantages of the *k-means* algorithm is that it requires prior knowledge of the number of sources. In this section it is assumed that the number of sources is known and *k-means* clustering is applied to the DOA data from (130) for each frame. The results for DOA estimates from the *k-means* algorithm are shown in Figure 68, where it can be seen that the estimates for all three sources are approximately equal when an average for all frames is found. When individual frames were analysed, a similar result was found.

There are several reasons why the *k-means* algorithm fails to give a meaningful result for DOA estimation, these include:

1) The algorithm assigns a centre for each cluster and assigns data points that are close to the midpoint of the cluster and these mid points are updated as more samples are added to the clusters. The output from the algorithm is the

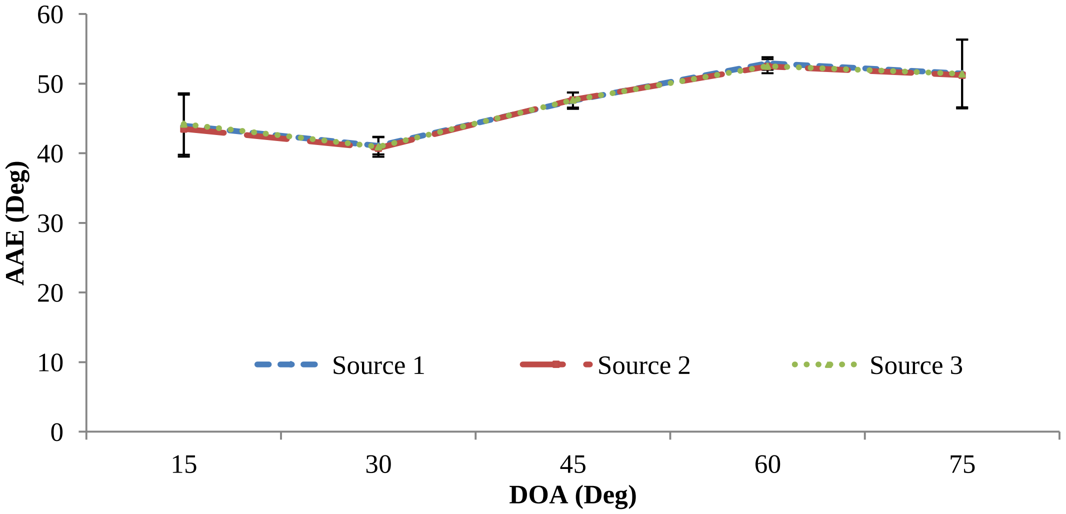


Figure 68 : The results of DOA estimates from the frequency domain intensity algorithm using *k*-means clustering for three sources (Error bars indicate 95% confidence intervals).

mean of the clusters which may not be the correct, as due to few data points in the cluster the mean of the cluster may move.

2) It does not eliminate those samples that are from reflections and due to errors
3) The accuracy of the algorithm depends on the number of clusters and without analysing the data there is no way of knowing how many clusters should be formed to get an accurate result.

From these results it can be shown that *k-means* algorithm in its original form cannot be applied for DOA data from (130).

4.6 Conclusions and Summary

The results presented in this section have shown that by taking advantage of the directional information from pressure gradient capsules in the AVS array an accurate estimate for DOAs can be obtained for a single source and for multiple sources. The results obtained for the DOA estimation with AVS and Soundfield microphone shows that AVS is capable of providing DOA estimates for stationary speech sources with AAE’s error’s of 1.58^0 while for the Soundfield the AAE is 4.99^0.

The accuracy of AVS is reduced for moving speech sources and the AAE increased from 1.58^0 to an average of 4.6^0 and similarly the error for the Soundfield microphone also increased for moving sources from 4.99^0 to 49.76^0. Although the error for moving sources has increased for the AVS, the error is less than 5^0. Further, the results show that AVS is capable of making accurate DOA estimates with frame sizes of 10 and 20 ms for moving sources.

In addition to the results for stationary and moving sources the work presented in this chapter has described a new technique for evaluating the DOA estimates in the frequency domain such that an accurate DOA estimate can be obtained for multiple sources. It has been shown that a direct averaging of all the DOA for each frame does not give an accurate DOA estimate and a clustering technique is needed to get an accurate DOA estimate for multiple sources.

Furthermore, it has been shown that for multiple sources the best method for obtaining DOA estimates is to use a frequency domain algorithm, as time domain algorithms fail to give as statistically valid estimate of the DOA for multiple speech sources. The AVS array is capable of providing DOAs for two and three sources with errors as small as 5 degrees, compared with the Soundfield microphone which produced error of 12 degrees for the two sources and 14 degrees for three sources.

The work in this chapter has shown that an AVS has the ability to give highly accurate DOA estimates in reverberant conditions for stationary speech sources; moving speech sources; and single and multiple speech sources. These results are significant as applications such as tracking moving sources and to obtain a DOA for multiple sources using a compact co-located microphone array. In the next chapter, methods for enhancing noise corrupted speech sources based on the AVS will be presented.

Chapter 5 Speech Enhancement with an AVS

5.1 Introduction

The work presented in the previous two chapters have shown that using an AVS array, an accurate DOA estimate for speech sources can be obtained under different scenarios. In this chapter, the recording from the AVS array will be used for enhancement of speech sources corrupted by diffuse noise and reverberation. In addition to the enhancement, a source separation technique that takes advantage of the directional information from the AVS array will be presented.

The enhancement of speech sources corrupted by diffuse noise and reverberation is extremely important for applications such as hands free telephony and video teleconferencing. There are several single channel algorithms that have been proposed for enhancement, such as Weiner filters and Kalman filtering, but in recent years it has been shown that by using multichannel recordings, much better improvements in terms of SNR can be obtained when compared to the single channel case [28].

In this chapter, three different techniques for enhancing speech sources corrupted by diffuse noise that take advantage of the directional recording of the AVS will be presented. These are speech enhancement based on beamforming; speech enhancement by perceptual filtering; and speech enhancement by using source separation technique.

The work presented here will show by applying conventional beamforming algorithms to the AVS array outputs, an improvement in PESQ MOS is obtained. Furthermore, it will be shown that by introducing a technique for obtaining a covariance matrix that represents the noise covariance matrix for the MVDR beamformer, further improvements in perceptual quality is obtained.

Weiner filters have been used in enhancement for several decades, and perceptually motivated wiener filters for single channel applications [89, 90, 154] have been shown to give good improvements in terms of perceptual quality. Here, a similar perceptual Wiener filter that is based on a multichannel scenario and takes full advantage of the directional characteristics of the AVS array will be presented. It will be shown that by using the recording of the AVS array, it is possible to get a closer match to the LP spectra of the speech signal that needs to be enhanced. Furthermore, different methods for obtaining multichannel LP spectra from the AVS array will also be discussed and the results for different methods will be compared.

BSS algorithms have been used for source separation, speech enhancement and DOA estimation. In this work, the fast ICA and convolutive fast ICA algorithms will be used for enhancement of noise corrupted sources. It will be shown that due to the directional characteristics of the AVS array, the fast ICA can be applied successfully to the AVS array outputs to obtain an enhanced signal. It will be shown that when compared to other arrays such as the Soundfield microphone, enhancement of the recording from the AVS array gives better results.

In addition to the enhancement of speech sources, a technique based on the directional information for the separation of mixed speech sources will be presented. Unlike most other BSS algorithms, the method presented here will be based on a co-located multichannel scenario. This is a very important distinction between the work presented in this chapter and other BSS algorithms, as one of key conditions for most BSS algorithms is that the channels used in the separation are from spatially distributed microphones.

The majority of BSS algorithm found in the literature are not algorithms that can be used for real time applications such as teleconferencing. In contrast, the algorithm that will be presented in this chapter for source separation will be able to perform source separation in real time. The results of the source separation algorithm will be compared against the well known ICA algorithm in terms of improvements in SIR, SDR, PESQ MOS tests and MOS listening tests with real listeners.

The effect of reverberation on speech signals is one of the most common problems in the enhancement. There are several algorithms that have been proposed to address this problem, but most of the algorithms that are proposed require the room impulse response to be known and in addition to the room impulse response most of these algorithms proposed for dereverberation are for single channels.

This chapter presents, a technique that does not rely on the room impulse response and takes advantage of directional characteristics of the AVS. It will be shown that this algorithm when applied to recordings made in a room with $RT_{60} > 1s$, there is a significant improvement in the processed recordings, furthermore the results of the proposed technique will be compared against the Multichannel Spatiotemporal

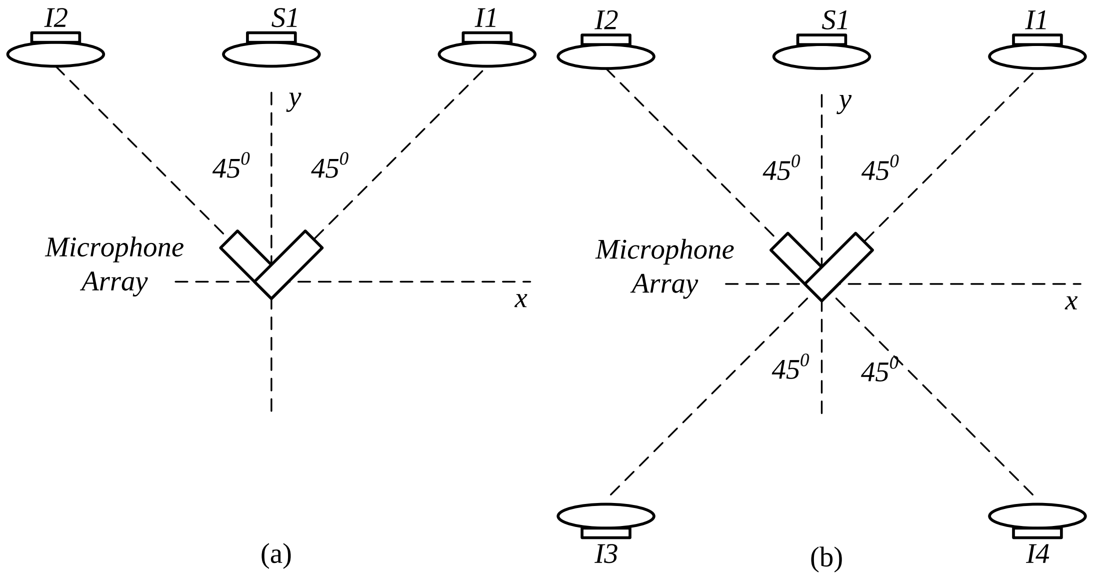


Figure 69 : Arrangement of Sources and Microphones for simulation and Experimental recording a) One source and two interferers b) One source in diffuse noise

Averaging Method for Enhancement of Reverberant Speech (SMERSH) algorithm [108].

The rest of this chapter will be organised as follows: a description of the experimental setup and the database created for evaluating the different enhancement algorithms will be presented in Section 5.2, followed by enhancement of noise corrupted speech source by beamforming methods in Section 5.3. In Section 5.4, enhancement work using the perceptually motivated enhancement algorithm will be presented followed by Section 5.5, where enhancement of the AVS outputs using fast ICA will be presented. In Section 5.6, a source separation algorithm for the AVS array will be presented and methods for obtaining accurate LP spectra for perceptual filtering will be presented in Section 5.7. An extension of the source separation algorithm of Section 5.6 will be used for dereverberation in Section 5.8 and finally, conclusions and summaries of the key results will be presented in Section 5.9.

5.2 The Experimental Setup and Database of Recordings

Experiments were performed to compare the performance of different enhancement algorithms for speech enhancement using simulated and real recordings from various types of microphone arrays in anechoic and reverberant conditions.

5.2.1 Experimental Setup for Real Recordings

Six female and six male speech sentences from IEEE speech corpus [139], each 10s long with 1s of silence at the start and at the end, were used as the test database. Noise sources include 10s segments of babble, recordings of a factory floor, recordings of the background noise of a moving vehicle, white noise and pink noise [155].

Two scenarios for sources are used a) one source, two interferer b) one source and diffuse noise (synthesized using four interferers), as shown in Figure 69 (a) and (b). Noisy speech signals were recorded with a range of SNR ranging from 0 dB to 20 dB (0dB – the signal and noise levels are equal) in increments of 5 dBs. Recordings were made at a sampling rate of 48 kHz and then down-sampled to 16 kHz before being processed by the enhancement methods. In total, one hundred recordings were made for each of five SNR levels. The recordings were made both in an anechoic chamber [22] and a room with a RT_{60} of 30ms.

5.2.2 Evaluation of Results

The enhanced speech signals were first analyzed using the ITU-PESQ software [115]. When using PESQ, each output from the enhancement approaches is compared with the original clean source signal to get a MOS for Listening Quality (MOS LQO) [115]. A difference MOS is generated by subtracting the MOS of an omni-directional recording of the mixed sources (used as the reference) from the MOS of the filtered outputs.

In addition to the PESQ, a MOS listening test of the filtered signals was carried out according to [114] in some experiments. The listening tests include twenty listeners, all native English speakers (ten male and ten female). Since the number of files and how the listening test were carried out for different experiments varied, a detailed description of the listening test for the specific experiments will be presented in the relevant sections. The results presented in this chapter include 95% confidence intervals.

5.3 Speech Enhancement Using Beamforming Techniques

The concept of beamforming and different types of beamformers has been discussed in detail in Chapter 2. Here, four different beamformers which were discussed in Chapter 2 will be applied to the AVS array for enhancing the outputs of noise corrupted speech source described in Section 5.2.1.

The four beamforming approaches for the AVS array that will be presented are:

1) Summing beamformer for AVS channels
2) The Griffiths and Jim beamformer
3) MVDR beamformer
4) Enhanced MVDR beamformer

5.3.1 The Compensation for Difference in Frequency Response of Different Microphone Capsules it the AVS

The output of the AVS array has been presented in Section 3.3, which were used for DOA estimation. Since the pressure gradient sensors produce a direct representation of the particle velocity as shown from (112), the frequency responses of these microphones are different to that of the omni-directional microphone which is a direct representation of the pressure at the array. The frequency responses of the two microphones are shown in Figure 20 and Figure 21, where it can be seen that the pressure gradient microphone has a high-pass effect. This high-pass effect can be assumed to be similar to the pre-emphasis filter which is required in applications such as linear prediction of speech. Hence, the pressure gradient sensors of the AVS can be assumed to introduce pre-emphasis like effect which will be confirmed in Section 5.4.3.

When using the output from the omni-directional sensor with the outputs from the gradient sensors, the output from the omni-directional sensor is pre-emphasised such that the three channels have a similar frequency response. The pre-emphasis is performed according to [156]:

$$o(t) = o(t) - 0.96 \times o(t-1) \tag{135}$$

After the processing of the AVS channels with the enhancements algorithms the outputs from these algorithms are de-emphasised according to [156]:

$$out(t) = out(t) + 0.96 \times out(t-1) \qquad (136)$$

where $out(t)$ is the output from the enhancement algorithm.

5.3.2 Summing Beamformer for AVS Channels

In the case of co-located microphone arrays like the AVS, the simplest beamformer is a summation of the channels. Unlike the ULA and spherical arrays where the microphone capsules are spatially located, due to which a time alignment of the signals are required, in the AVS the microphone capsules are co-located hence a simple summing of the channels can be performed. The AVS summing beamformer can be expressed as:

$$B(t,\theta) = \beta_0 o(t,\theta) + x(t,\theta) + y(t,\theta) \qquad (137)$$

where β_0 is either 1 or 0 and switches on/off the omni-directional component $o(t,\theta)$ and $o(t,\theta)$, $x(t,\theta)$ and $y(t,\theta)$ are defined in Section 3.4. It will be shown later in this chapter that due to level of noise captured by the omni-directional sensor, by excluding it in the beamformer as described above, a better outcome can be achieved.

5.3.3 The Griffiths and Jim Beamformer

The Beamformer proposed by Griffiths and Jim (GJ) (also known as the Generalised Sidelobe Canceller (GSC)) was discussed in Section 2.9.11 which is an improvement to the LCMV beamformer. As described before, the advantage offered by the GSC algorithm is that it offers a data independent solution to the LCMV beamformer and it provides a mechanism for changing a constrained minimization problem into an unconstrained form. The basic idea proposed in the GSC algorithm is to divide the filter of the LCMV method into two components operating in orthogonal subspaces. As described in Section 2.9.13, one component is the fixed beamformer, which in the case of the AVS is the beamformer described by (137) in the previous section. The other component is the blocking matrix which rejects the desired signal and an adaptive filter as explained in Section 2.9.13.

One of the drawbacks of this beamformer is leaking of the signal from the blocking matrix, and several solution have been proposed to limit the signal leaking

[82]. Here, the improved version of the GJ beamformer described in [82] is implemented for beamforming the AVS outputs.

5.3.4 The MVDR Beamformer

The MVDR beamformer used in this work is based on the frequency domain version proposed in [128]. The MVDR Beamformer forms a filter $\boldsymbol{w}$ which minimizes the output power without introducing any distortions [69]:

$$E[Z_k^2] = \boldsymbol{w}^T\boldsymbol{R}\boldsymbol{w} \tag{138}$$

where $\boldsymbol{R}$ is the covariance matrix in the frequency domain. The implementation of the beamformer is as follows; An FFT of size 1024 is found using a hamming window with an overlap of 50 %. The sample matrix in the frequency domain is represented as:

$$\boldsymbol{X}_k = [o_k \quad x_k \quad y_k] \tag{139}$$

where n is the frame number and k is the frequency bin. The $F = 32$ most recent frames are buffered and the covariance matrix $\boldsymbol{R}$ of $\boldsymbol{X_k}$ is found according to [128].

$$\boldsymbol{R}_k = \begin{bmatrix} \frac{M}{F}\sum_i^F o_{ki}^* o_{ki} & \frac{1}{F}\sum_i^F o_{ki}^* x_{ki} & \frac{1}{F}\sum_i^F o_{ki}^* y_{ki} \\ \frac{1}{F}\sum_i^F x_{ki}^* o_{ki} & \frac{M}{F}\sum_i^F x_{ki}^* x_{ki} & \frac{1}{F}\sum_i^F x_{ki}^* y_{ki} \\ \frac{1}{F}\sum_i^F y_{ki}^* o_{ki} & \frac{1}{F}\sum_i^F y_{ki}^* x_{ki} & \frac{M}{F}\sum_i^F y_{ki}^* y_{ki} \end{bmatrix} \tag{140}$$

where $M = 1.03$ which is regularization constant to help avoid matrix singularity and $*$ is complex conjugate. The covariance matrix is updated every 16 frames. The MVDR filter is expressed as [128]:

$$\boldsymbol{w}_k = \frac{\boldsymbol{R}_k^{-1}\boldsymbol{h}}{\boldsymbol{h}^T\boldsymbol{R}_k^{-1}\boldsymbol{h}} \tag{141}$$

where $\boldsymbol{h(\theta)} = [1 \quad cos\theta \quad sin\theta]$ is the steering vector for an AVS [9] and with the optimizing constraints for each frequency band given as:

$$\min_{\boldsymbol{w}} \boldsymbol{w}^T\boldsymbol{R}_{ky}\boldsymbol{w} \quad \text{subject to} \quad \boldsymbol{w}^T\,\boldsymbol{h} = 1 \tag{142}$$

The output of the beamformer for each frequency band k is given by:

$$Z_k = \boldsymbol{w}_k^T\boldsymbol{X}_k \tag{143}$$

The time domain output is obtained by using the inverse FFT and performing an overlap add of the frames. Here, the minimization of the filter is based on the covariance matrix of the AVS output channels. The idea of the minimization is to reduce the interference and noise components. Hence, the covariance matrix of interferers and noise has to be used to get the best performance from the MVDR beamformer [77] . The problem with getting the covariance matrix of interferers and noise is that in real applications these matrices are not available [77]. Hence, to get a better estimate of the covariance matrix, a solution is provided in the next section.

5.3.5 Enhanced MVDR Beamformer

The improvement proposed in this section is based on an SVD approach applied to the covariance matrix estimation used in the MVDR Beamformer described in the previous section. A similar approach was proposed based on the Eigen decomposition of the covariance matrix in [157] where the noise components from the Eigen decomposition were filtered such that only the source and interferer were used in the formation of the covariance matrix, here in contrast to the approach of [157] the covariance matrix is formed from the noise and interferers only. As described in the previous section and in Section 2.9.11, the MVDR Beamformer is derived on the assumption that the covariance matrix of the array output is a close match to that of the covariance matrix of the interferer and noise [70]. The method proposed in this section is an improvement which estimates the interferer and the noise components in the array output using SVD.

The equations describing the outputs of the AVS are given in (116-119) from which it is seen that the outputs of the AVS contain the source as well as the undesired noise. Hence, performing SVD will result in an estimate of the noise, as well as the source signal in the channels.

To get an accurate noise estimate the AVS outputs are paired, such that $o(t)$ is paired with $x(t)$ and $o(t)$ is paired with $y(t)$, to form two vectors $X\ (t)$ and $Y\ (t)$, as shown below:

$$X(t) = [o(t) \quad x(t)]^T \tag{144}$$

$$Y(t) = [o(t) \quad y(t)]^T \tag{145}$$

where T is transpose and each of these matrices are $M \times N$, where $M\ =\ 2$ and N is the number of samples. The SVD of matrix $X(t)$ is expressed as:

$$X(t) = USV^* \tag{146}$$

where U is a $M \times N$ with orthonormal columns ($U^T U = I$) where I is the Identity matrix, where V is a $N \times N$ orthonormal matrix and S is a $N \times N$ with diagonal positive or zeros values called the Singular matrix and square of the diagonal elements are the Eigen values of the matrix $X(t)$. The smallest eigen values of the matrix $X(t)$

corresponds to the noise [58]. Similarly, the SVD is performed on the $Y(t)$. A new matrix $\widehat{XY}(t)$, $M \times N$ large is formed from the smallest values of S from each of the SVD operations.

This process effectively creates a matrix that contains noise components from (116-119) of the AVS output and reduces the three channels of the AVS to two channels. The covariance matrix R in (140) is now formed from $\widehat{XY}(t)$ and is used in the MVDR beamformer from the previous section.

5.3.6 The Results of Applying the Beamformers to the AVS Outputs

The enhanced speech signals were analyzed as described in Section 5.2.2. The results of the experiments shown in Figure 70 and Figure 71 are for average difference MOS (difference MOS is the difference between the MOS of clean omni-directional recording and the MOS for the output of the enhancement algorithm) for AVS outputs of different types of diffuse noise and for averaged SNR, for a target at 45 degrees in azimuth, filtered with different beamforming algorithms in anechoic and reverberant conditions.

The results show that the proposed method for estimating the noise and interference covariance matrices does offer advantages over the conventional use of the covariance matrix of the array output. This is seen from the results of the MVDR beamformer and the enhanced version of the MVDR Beamformer where an improvement of 0.4 and 0.3 MOS is obtained in anechoic and reverberant conditions, respectively. Furthermore, the results also show that the proposed enhancement to the MVDR Beamformer works best with noise types, pink, white, moving vehicle and factory.

In comparison the GJ beamformer has shown better performance compared to the original MVDR implementation with an improvement in MOS of 0.3 and 0.1 in anechoic and reverberant conditions. Furthermore, it is seen from the results that all algorithms perform better in anechoic conditions. From the results it is also clear that as the SNR increases the performance of the beamformers are reduced.

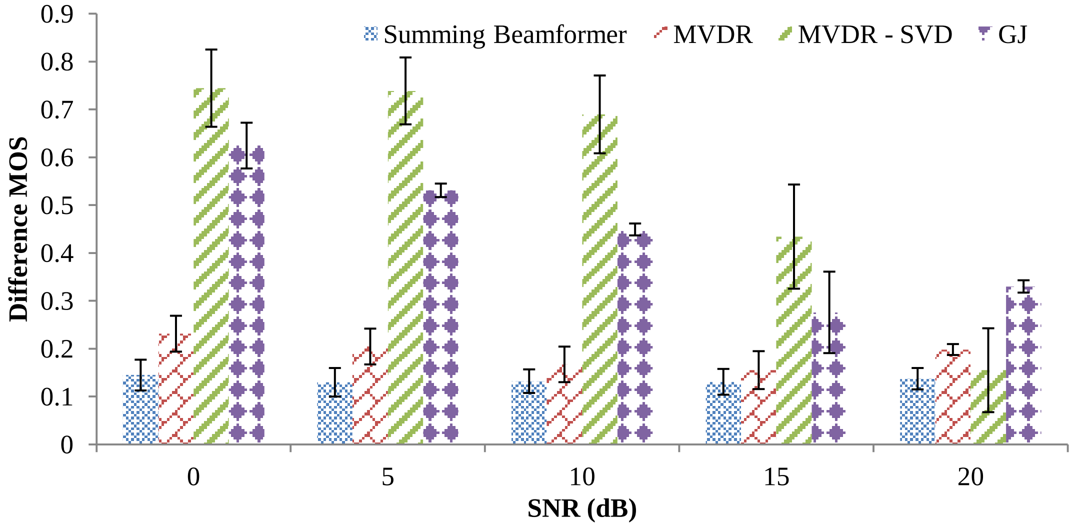


Figure 70: Results for Difference MOS LQO for different beamformers for recordings in anechoic conditions Error bars indicate 95% confidence intervals).

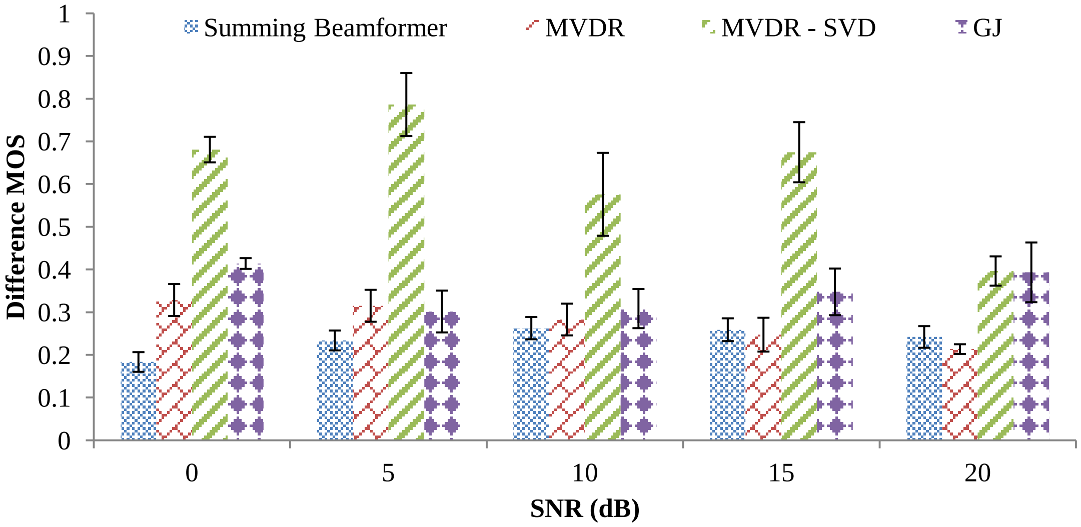


Figure 71: Results for Difference MOS for different beamformers for recordings in reverberant conditions Error bars indicate 95% confidence intervals).

5.3.7 Results of Listening Test for Different Beamformers

The results presented in this section and shown in Figure 72 are for listing tests carried out for different beamformers according to [114]. The listening tests include twenty listeners, all native English speakers (ten male and ten female) and the listening tests were carried for all different types of noise. The test contained six files for each

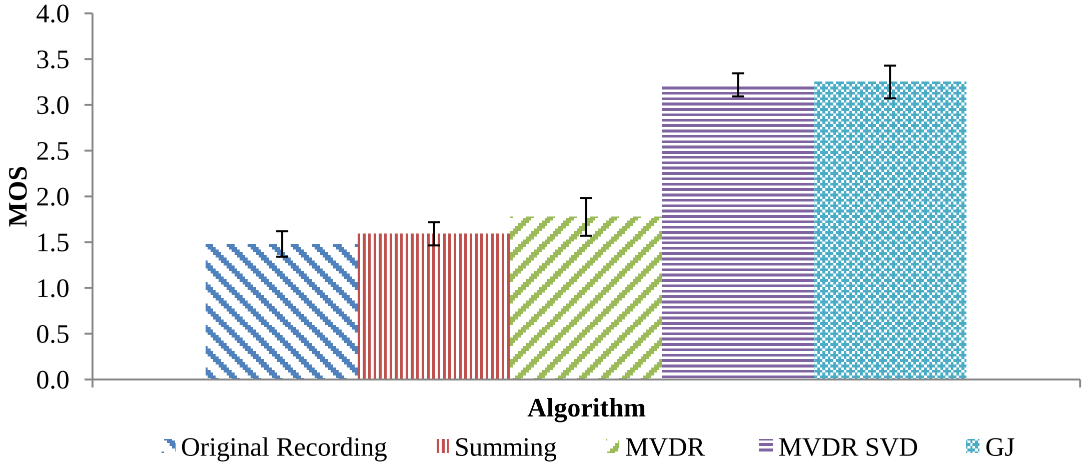


Figure 72: The results for listening tests for different beamformers (Error bars indicate 95% confidence intervals).

type of beamformer, which is randomised. The files tested included files recorded in anechoic and reverberant conditions, and unprocessed files.

The results show that the best beamformers are GJ and enhanced MVDR beamformer which scored MOS score of 3.3 and 3.2 respectively and the unprocessed files scored 1.5 MOS. This is an improvement from bad to fair on the MOS scale of Table 1. The results for the MVDR, Summing and Original recording all scored approximately equal MOS results. Although the results from the listening test show a similar pattern to that of the PESQ results, difference MOS results for the listening test were generally higher than for the PESQ results.

5.3.8 Summary

In this section, four different methods for beamforming the outputs of the AVS array has been presented. The performance of these beamformers has been evaluated using subjective and objective perceptual tests. The results of these tests show that in terms of the enhancement the enhanced MVDR and GJ beamformer performed the best. The result presented in this section has shown that by modifying the MVDR beamformer as proposed, the performance of the MVDR beamformer improved significantly. The next section in this chapter will look at multichannel perceptual filtering.

5.4 Linear Predictive perceptual Filtering for Acoustic Vector Sensors: Exploiting Directional Recordings for High Quality Speech Enhancement

A fundamental stage of most speech coders is the LP spectrum estimation. In noisy environments, degradation in signal quality leads to inaccurate estimation of the LP spectrum and hence reduces the speech coding quality, such as used in hands free communication using mobile phones. A typical solution to this problem is speech enhancement of the recorded signal prior to speech coding. Speech enhancement using microphone arrays offers superior performance over a single microphone in reducing both speech signal distortion and speech intelligibility degradation resulting from noise removal [28].

In this section, the outputs from the AVS are exploited within a speech enhancement technique that combines beamforming and LP spectrum based perceptual filtering. The use of gradient sensors allows for precise recording of directional sound and minimization of the effects of both diffuse noise and reverberation [5] and these hardware advantages enable improved accuracy in estimating the LP spectrum in noisy environments.

In [90], postfilters based on LP spectral models, typically used in speech coding [156], were applied to the problem of enhancing single channel speech. Recently, an approach to speech de-reverberation based on an LP-based postfiltering approach for 2 channels of a circular microphone array reported good results in terms of perceptual quality improvement [154]. In this section the technique of [90] is adapted for the AVS and the results presented demonstrate improved performance in LP modelling of speech spectrum compared to single channel approach of [90]. Here, subjective and objective speech quality results are also presented and show significant improvements compared with an existing speech enhancement technique for the AVS based on the Minimum MVDR beamformer [128].

5.4.1 Perceptual LP Filtered Beamforming Using an AVS

The proposed speech enhancement system shown in Figure 73 is composed of two main stages. Firstly, the AVS signals are combined to form a beamformed

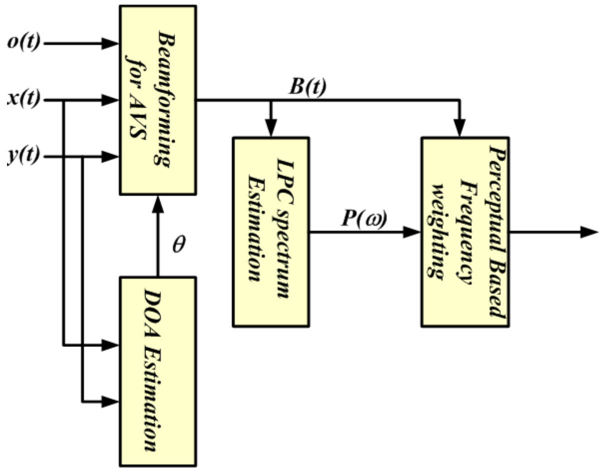


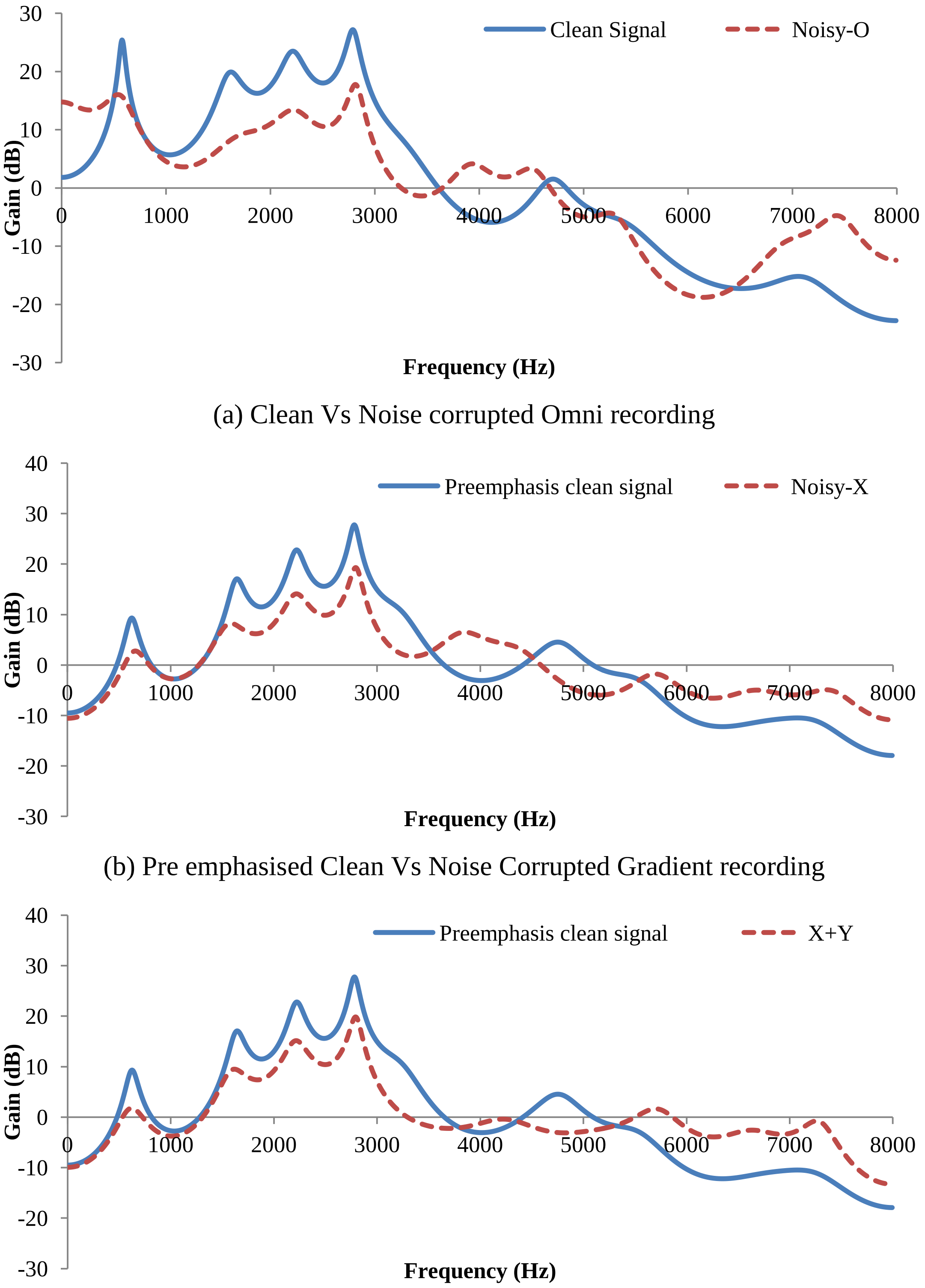
Figure 73: Block Diagram of the proposed system.

recording of the source, and secondly, the beamformer output is fed to a perceptually adaptive frequency weighting filter. This filter is based on the LP spectra of the gradient signals derived from the beamformer output.

5.4.2 The DOA Estimation and Beamforming Stage

The beamforming stage in the block diagram of Figure 73 is a crucial part in the performance of the proposed algorithm. The beamformer combines the AVS channels such that a more accurate estimate of the LP spectra of the speech in the current frame can be obtained. The performance of the algorithm depends on the accuracy of the beamformer output. In Section 5.3, several beamforming techniques for the AVS array has been presented. In this section the summing beamformer will be used initially. A study on the effect on the using a more complex beamformers and other methods for combining the output channels of the AVS will be presented later in this chapter.

The DOA estimation Block is needed if the beamforming algorithm used is more complex algorithm such as the MVDR beamformer, which requires the DOA estimates, but here, since the beamformer is a simple summing operation, the DOA estimation block can be ignored.



(a) Clean Vs Noise corrupted Omni recording

(b) Pre emphasised Clean Vs Noise Corrupted Gradient recording

(c) Pre emphasised Clean Vs Output of Summing beamformer for Gradient sensors

Figure 74: LP spectrums of vowel ‘a’ at Azimuth 45^0 and 0 SNR.

5.4.3 Enhancement of LP Spectra of a Noisy Speech Signal

The LP spectra of the noise corrupted speech signals are shown in Figure 74 for different microphones capsules of the AVS. The LP spectra of the omni-direction sensor and the clean speech is presented in Figure 74 (a), where it can be seen that the formant peaks of the LP spectra are not defined, especially in the regions that are most important for speech between the 0 and 4 kHz.

In Section 5.3.1 the difference in frequency response between the pressure sensors and the pressure gradient sensor were discussed. It was proposed that in order to compensate for the high-pass effect of the pressure gradient sensors the output of the pressure sensors have to be pre-emphasised.

The pre-emphasised clean signal and the output from the pressure gradient sensors are shown in Figure 74 (b). From Figure 74 (a) and (b) it can be seen that the pre-emphasised version of the clean source is a closer match to the recording from the pressure gradient sensor. From Figure 74 it can be seen that the recordings of the noise corrupted pressure gradient sensors show a much closer match to the clean recordings compared to the omni-directional recordings. The formant peaks of the spectra of the pressure gradient sensor are much more defined, especially in the frequency regions of 0 to 4 kHz which is critical for speech.

Since the individual pressure gradient sensors have a close match to that of the clean speech, beamforming the outputs of the pressure gradient sensor will give an even closer match to the clean signal. In Figure 74 (c) the LP spectra for the output of the summing beamformer in (137), with $\beta = 0$ is presented. From the plot it can be seen that a closer match to the LP spectra of the clean source is achieved, and it is seen that there is an improvement over the single channel case.

The plots of Figure 74 show that the LP spectra of the output of the pressure gradient sensors show a closer match to that of the clean signal, but it does not give a measure of exactly how close the two LP spectra are. There are several methods for evaluating how close two LP spectra are, of which the two most commonly used methods are:

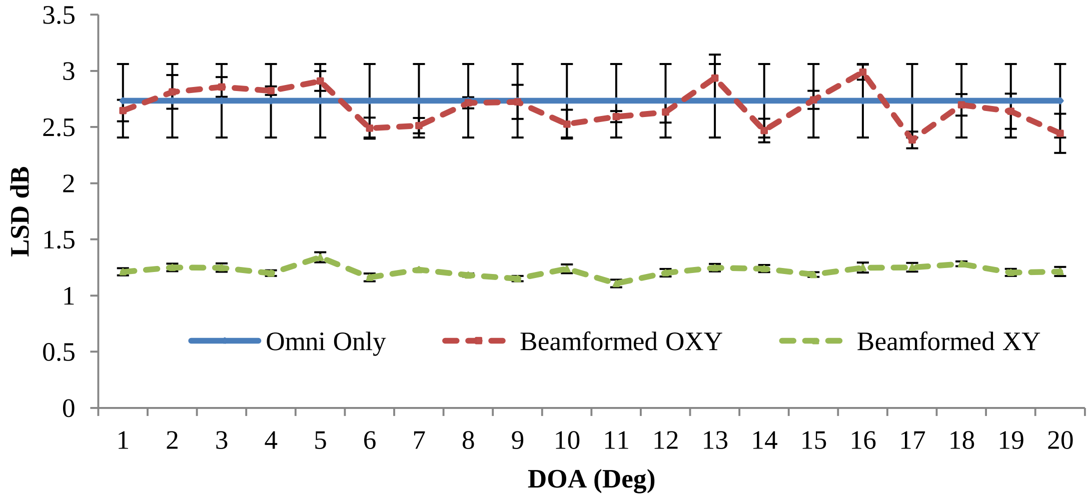


Figure 75: LSD for Beamformer output (Error bars indicate 95% confidence intervals).

1) The Log Spectral Distortion (LSD)
2) The Saito Itakura Distance (SID)

which were discussed in detail in Chapter 2. Figure 75 shows the LSD [158] between the clean signal and the calculated LP spectra for beamformed gradient components, beamformed omni and gradient components and the omni-directional component. The LSD results were generated using the database and the recording setup of described in Section 5.2.1, where in total fifty recordings are made for each source azimuth between 0 and 20 degrees (in steps of 5 degrees). The LSD results show how close the two spectra are, as the LSD increases the difference between the two spectra increases. The plots in Figure 75 show that when the omni-directional sensor is used, on average, the LSD is a relatively constant at 2.7dB compared to 1.2dB when only the gradient sensors are used. This result confirms what has been shown in Figure 74. Hence, when forming the LP spectra by excluding the omni-directional sensor, a much better estimate of the LP spectra can be obtained.

This difference in gain of the LSD measures is caused by the directionality of the gradient microphones, which reduces the amount of degradation caused by noise and reverberation. The effect of directionality can be seen by analysing the amount of noise captured by the directional microphone in a diffuse noise field.

Figure 76 shows the SNR for the x and y gradient sensors of the AVS in diffuse noise. Recordings are made of 12 speech sentences (six male and six female) in diffuse white

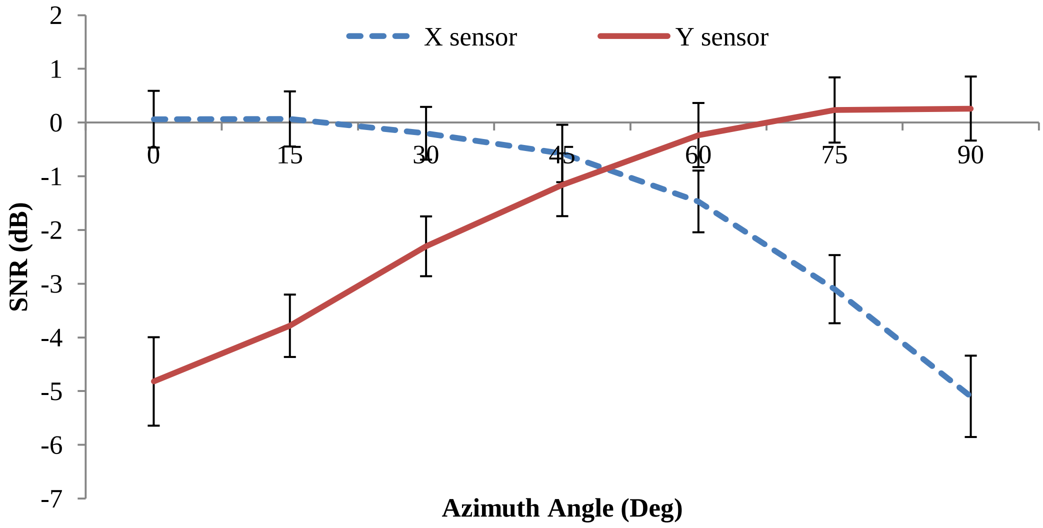


Figure 76: The SNR for the *x* and *y* channels at different DOAs (Error bars indicate 95% confidence intervals).

and pink noise with SNR values at 0dB and 10dB (0dB noise energy and signal energies are equal). SNR values are calculated as described in Section 2.9.

The results present in Figure 76 are for the SNR for a source at 0 degrees in azimuth and recordings are by rotating the AVS in azimuth in increments of 15 degrees from 0 to 90 degrees. As the array is rotated in azimuth the level of noise on the sensor that is parallel to the target remains constant while the level of target signal increases. On the other hand, the level of target signal on the sensor perpendicular to the target reduces while noise levels remain constant. It can be seen from Figure 76 that the level of target on both sensors are approximately equal at around 45^0. From these observations it can be concluded that for azimuth angles approximately from 0-20 degrees and 70-90 degrees the recordings of one sensor has dominant background noise and the other sensor has a dominant target signal in diffuse noise.

These results show that for the AVS if a source is at 90 degrees or when the source is at 0 degrees, then the perpendicular component will contain just the background noise and this highlights the need for beamforming the x and y components such that the best possible LPC filter can be obtained by reducing the influence of the errors and noise.

5.4.4 The Perceptual LP Filter for an AVS

This system achieves speech enhancement by employing a perceptually motivated frequency based filter similar to the perceptual based wiener filter described in Section 2.9. The perceptual filter is similar to [90] and [154], which is conceptually based on the perceptual masking filter used in low rate speech coders such as CELP [156]. In speech coding, a perceptual weighting filter is employed during excitation vector search. The filter emphasises audible quantization noise in spectral valleys while noise near formant frequencies, which is masked, is de-emphasised. This results in a shift of quantisation noise in the decoded speech to the masked areas around formants. In [90], this approach is adjusted for speech enhancement in the frequency domain such that the noise from speech enhancement is minimised in the spectral valleys. For each frame the method can be described by:

$$\hat{B}(\omega) = g(\omega)B(\omega) \tag{144}$$

where, $B(\omega)$ and $\hat{B}(\omega)$ represent the recorded and enhanced speech spectra of (137), respectively, and $g(\omega)$ are the frequency coefficients of a modified version of a standard Wiener filter (87), whereby a frequency weighting is used during noise spectral subtraction based on a standard LP error shaping filter [158]. This shaping is controlled by the estimated SNR for each frame. The SNR is estimated for each frequency as the ratio of the estimated source signal (difference between the recorded and the noise estimate) to the noise estimate. In this work, the noise power is fixed and is calculated from the first 500ms of each recording (where speech is not present). In practice, a VAD can also be used to update the noise spectrum estimation. Here, spectrums are estimated for 20 ms frames with 50% overlapping Hamming windows and an LP order of 18.

The key difference in the application of the algorithm in [90] and described in Section 2.9 to this work is the threshold in the calculation of the SNR. In [90] there is an upper and lower bound for the SNR which is used in the calculation of $g(\omega)$, while in this work the value is adaptively controlled by updating the SNR for each frame. Informal listening tests found that this reduces musical distortions in the output of the speech enhancer where there are significant sudden changes to the amplitude of the speaker’s voice and hence sudden changes in SNR values.

5.4.5 Results of Applying the Proposed Filter

The evaluation of the results was performed using the ITU-PESQ software as described in Section 5.2.2, In addition to the PESQ a MOS listening test of the filtered signals was carried out according to [114]. The listening tests include twenty listeners all native English speakers (ten male and ten female). The listening tests contained three types of diffuse noise: babble; moving vehicle; and pink noise. Each test contained three types of files: files from the output of the proposed technique; files from MVDR Beamformer; and unprocessed files in both reverberant and anechoic conditions combined. Each listener was asked to listen to two sets of files from the 3 types of diffuse noise and each set contained 36 randomised files.

5.4.6 Simulation of the AVS Recording in Anechoic Conditions

The results presented and shown in Figure 77 are for the simulation of the AVS for an anechoic room. The simulation is carried out using Roomsim [159, 160] in matlab. The gradient sensor used in this simulation is modelled based on the actual gradient sensors used in the construction of the AVS. The simulation is carried at 0dB SNR with AVS rotated in azimuth from 0 to 90 degrees in steps of 5 degrees. For each step, 12 recordings are made (six male and six female sentences). Simulations are carried out for 2 diffuse noise conditions, white and pink noise.

The results are for average difference MOS of the filtered samples from the AVS. The results show there is an improvement of 0.66 difference MOS with the XY components filtered using the LPC spectrum formed from the XY components. The results also show that when the omni-directional component is added, the difference MOS drops to 0.56 for the sum of the OXY components filtered by LPC spectrum of the XY and sum of OXY filtered using the Omni. In Section 5.4.3 the effect of noise on the LP spectra was described. Here from Figure 77, the effect on the output of the filter due to the increased noise in the LP spectrum can be seen in plots of the OXY-LPC-XY and OXY-LPC-O. These results show, when the omni-directional component of the recordings is excluded in the beamformer, the results improved.

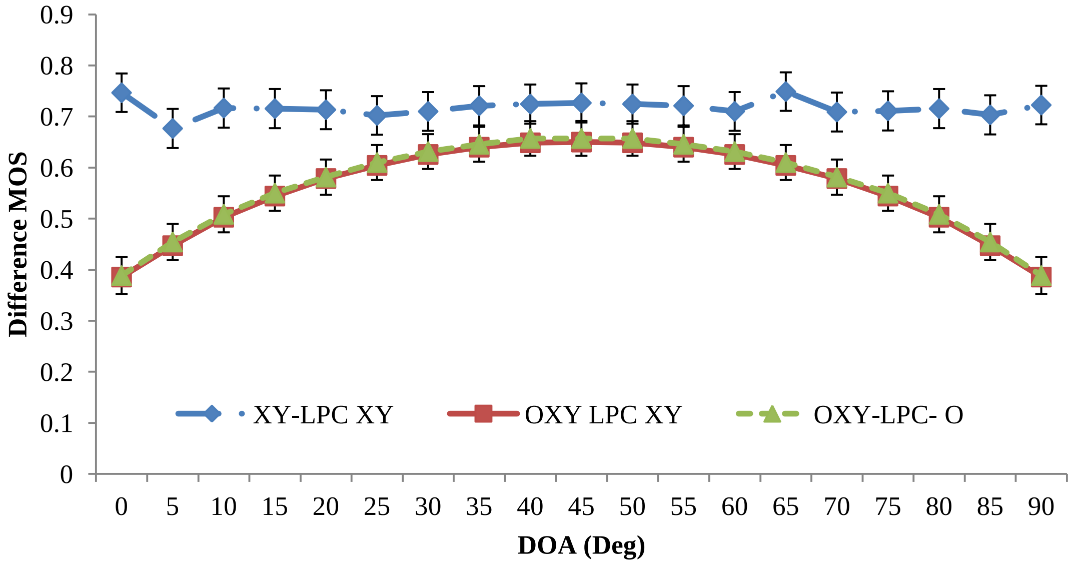


Figure 77: Difference MOS Results for simulated recordings. (Error bars indicate 95% confidence intervals).

XY-LPC-XY -the beamformer output of XY filtered with LPC spectrum of the beamformer output of XY.

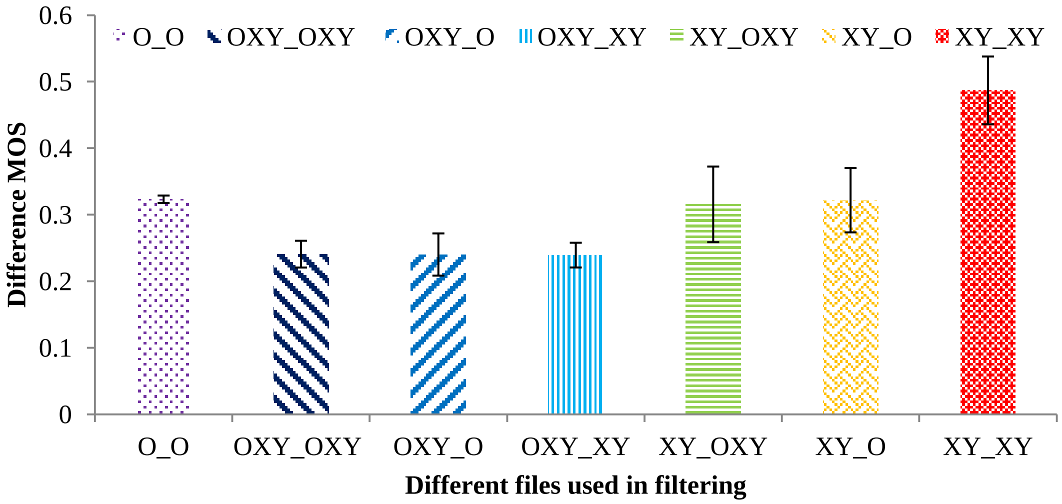
OXY-LPC-XY - the beamformer output of OXY filtered with LPC spectrum of beamformer output of XY.

OXY – LPC- O - the beamformer output of OXY filtered with LPC spectrum of O.

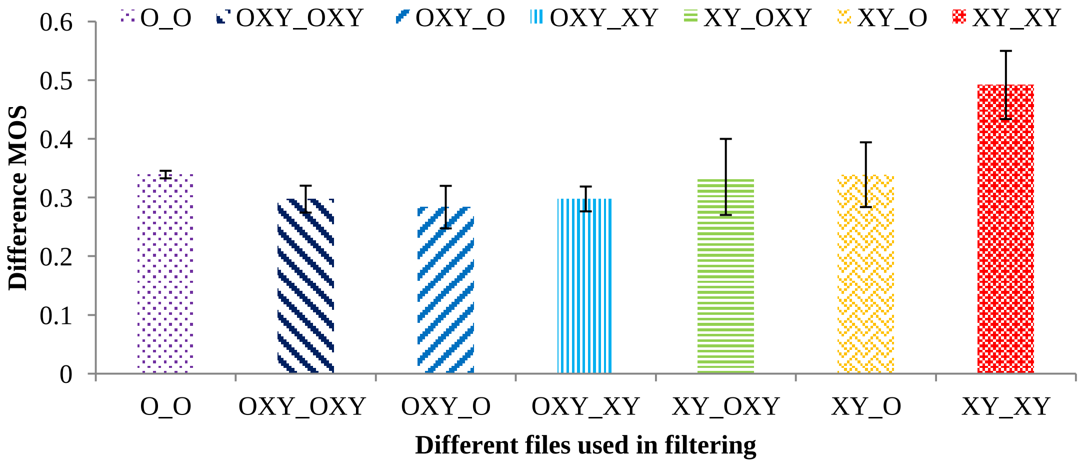
5.4.7 Experiments with Real Recordings

The results of the experiments shown in Figure 78 are for average difference MOS for AVS outputs of different types of diffuse noise and for averaged SNR, for a target at 45^0 in azimuth, filtered with different combinations of AVS outputs using the perceptual filter described in this section.

The results show that there is an improvement of 0.2 in terms of the difference MOS gained by leaving out the omni-directional sensor in LP spectrum calculation. As described in Section 5.4.3, the LP spectra of the omni-directional sensor are not as accurate as that of the gradient pressure sensors, and as a result the performance of the enhancer is reduced when the omni-directional sensor is included in the LP spectra calculation.



a) Anechoic Conditions



b) Reverberant Conditions

Figure 78: Difference MOS for output for different combinations of AVS outputs with the proposed method (Error bars indicate 95% confidence intervals).

O_O – omni is used in the LP spectra and filtering is done on O.

OXY_OXY – beamformed o, x and y is used in LP spectra calculation and filtering is done on beamformed o, x and y.

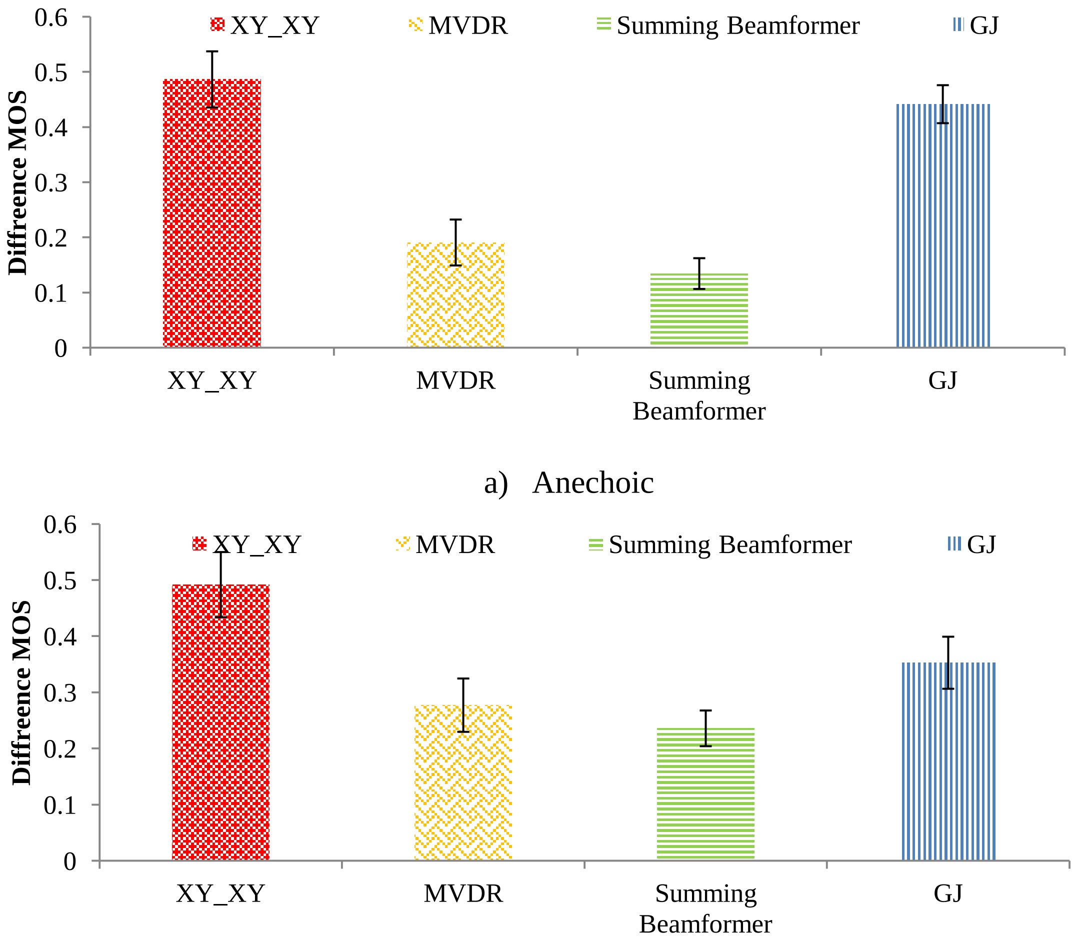
In addition, when the omni-directional sensor is used in the beamformer for the input of the filter, there is a drop of 0.2 difference MOS. The amount of noise in both anechoic and reverberant recordings of the omni-directional sensor and secondary reflection in the case of the reverberant recordings contribute to the poor performance of the system when the omni-directional sensor is included in the beamformer for the input of the filter. A similar result is obtained when including the omni-directional sensor in both the beamformer for input to the filter and the LP spectrum calculation.

The comparison of the proposed method with beamforming algorithms are shown in Figure 79. When compared against an MVDR beamformer, the proposed technique shows an improvement in difference MOS of 0.3 in anechoic and 0.2 in reverberant conditions. When compared with the Summing beamformer (without the perceptual weighting filter) the proposed technique shows an improvement of 0.4 differences MOS for anechoic and 0.3 difference MOS reverberant case. The performance of the GJ beamformer in anechoic conditions is very close to that of the perceptual LPC based filter; it can be seen from the results that the GJ beamformer has a difference MOS of 0.4, while the proposed technique has a difference MOS of 0.5, which is 0.1 improvement. However, but in reverberant conditions the performance of the proposed algorithm produces a improvement of 0.14 difference MOS, as the performance of GJ beamformer suffers in reverberant conditions. The results presented above show that a perceptually motivated multichannel wiener filter performs better than well known beamformers in terms of enhancing noise corrupted speech signals captured by an AVS.

Here, the MVDR beamformer has shown that there is very little improvement in the difference MOS for noise corrupted speech, but in [28], it was shown that the MVDR beamformer is in fact equivalent to a multichannel Weiner filter. Hence it can be said that the results also show a comparison between a multichannel Weiner filter applied to the outputs of an AVS against a perceptual multichannel Wiener filter.

Having established that the new technique is the best performing technique, compared with other algorithms tested, experiments were conducted to evaluate the impact of array orientation (relative to the source) on the resulting performance of the new technique. The signal to noise ratio is set at 0dB (the worst case scenario) and to 10dB while the recordings are made for all the speech sentences corrupted by white and pink noise. The arrays are rotated in azimuth through 90 degrees at 15 degree intervals and recordings made for each orientation.

The results in Figure 80 show averaged MOS for outputs from the new technique and the MVDR performed on these recordings in both anechoic and reverberant



a) Anechoic

b) Reverberant Conditions

Figure 79: Comparison of the proposed method with Different Beamformers (Error bars indicate 95% confidence intervals).

environments. The results show that there is very little or no effect on the performance of new technique or the MVDR by turning the array in azimuth.

5.4.8 Listening Tests for the Proposed Filter Outputs

Listening test were carried out as described in Section 5.4.5. The results of the listening test are presented in Figure 81 and show that on average for all the three types of noise there is an improvement in the difference MOS of 1.6 over the unprocessed recordings for the proposed technique and an improvement of 0.1 in difference MOS for the MVDR Beamformer over the unprocessed recordings.

The results also demonstrate that the diffuse noise sources babble and pink noise are removed more efficiently than that of the vehicle noise. The vehicle noise can be described as a low frequency hum which is in the range of speech, especially male. The

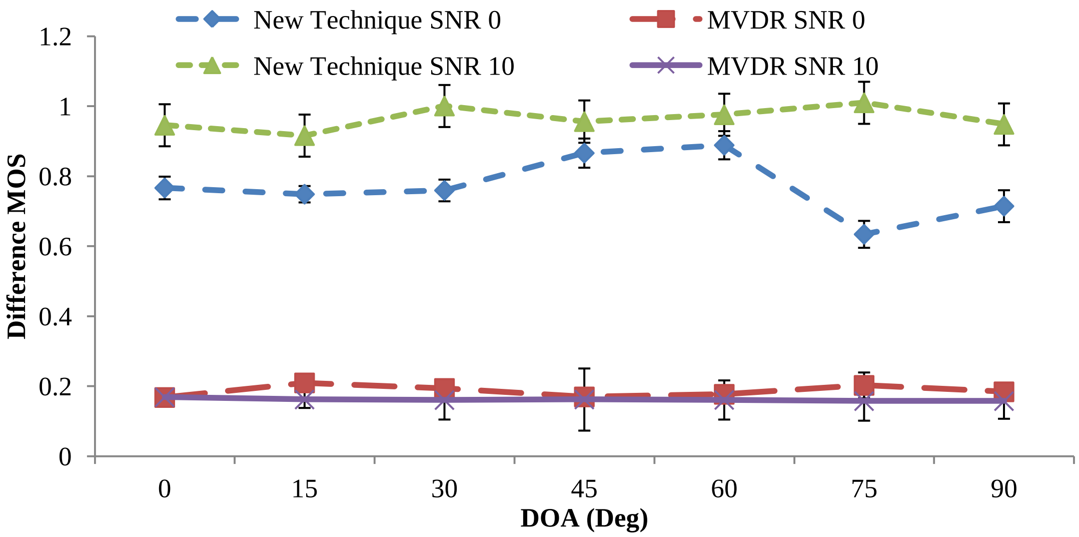


Figure 80: Difference MOS for Different Azimuth angles (Error bars indicate 95% confidence intervals).

mechanism used to filter out the noise in the technique described is to assume that the noise in the spectral valleys is filtered and noise close to the formant peaks is masked. When the noise is in the same frequency regions as that of the speech signals, formant peaks are unable to mask the noise.

There is a difference in the MOS score between the PESQ and the listening test. Although the relative performance measured by the objective and subjective tests are similar, the listening tests show that the improvement in the quality of the output from the new technique is rated 4 times higher by listeners than PESQ. This agrees with [114] where PESQ as a reliable estimate of subjective quality has not been completely validated for distortions such as caused by simultaneous talkers and artifacts from noise reduction algorithms. This highlights the importance of listening tests by real listeners to evaluate the actual quality improvement of speech enhancement algorithms.

5.4.9 Summary

This section investigated the use of the directional components of the AVS to improve the perceptual quality of speech in noisy environments. The proposed techniques use the gradient components of the AVS to generate LP spectra which is used to filter noise in the beamformer output, which improves the perceptual quality compared to existing state of the art algorithms such as the MVDR Beamformer. The results presented in this section shows that there is an improvement of accuracy in terms

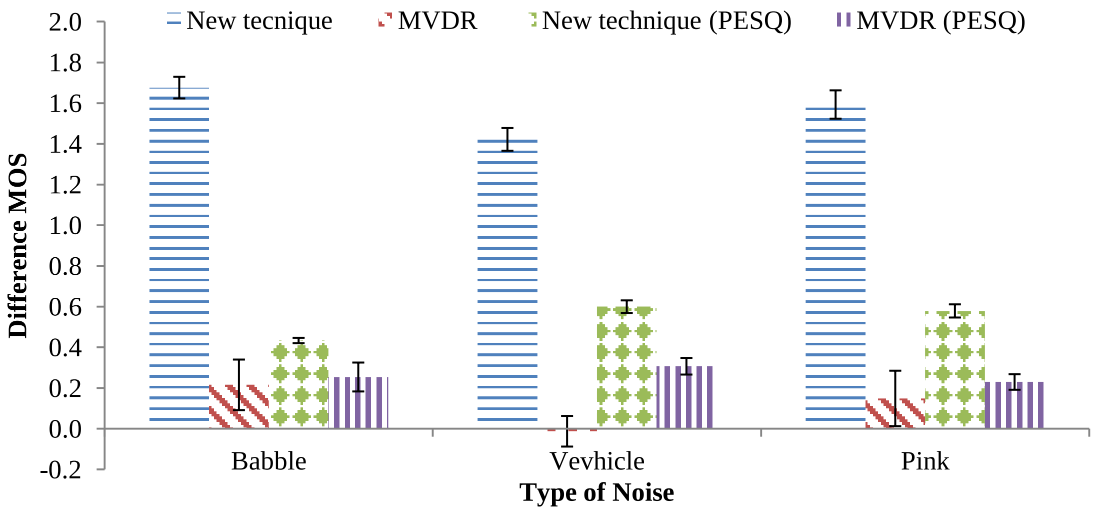


Figure 81: Difference MOS for output of the Filter from listening (Error bars indicate 95% confidence intervals).

of LSD of the LP spectra generated from the beamformed gradient components of the AVS compared to with using the omni-directional recording in the beamformer.

The work in this section shows a significant improvement in perceptual quality (measured using PESQ and listening tests) resulting from the proposed speech enhancement technique compared to other beamforming approaches applied to an AVS in both anechoic and reverberant environments. A key factor in the performance improvement is the use of directional recording using the AVS, which results in a more accurate estimate of the LP spectrum compared to that of a single channel omni-directional sensor.

Here, only the summing beamformer was used in the processes of obtaining the LP spectra and filtering by the proposed method, but from these results other beamforming techniques without the perceptual filter does produce improvements in perceptual quality of the noise corrupted speech. Hence, it is expected that by using more complex beamforming and speech enhancement techniques to obtain a LP spectra can be used with the perceptual filter to obtain further enhancement to noise corrupted speech. The use of more complex algorithms for estimation of LP spectra will be presented later in this chapter.

Beamforming algorithms and multichannel perceptual filtering for enhancement of speech sources has been presented in this chapter. Another approach for enhancement of speech sources is the use of source separation techniques. In the next section the fast ICA algorithm will be applied to the outputs of the AVS array for the same database of noise corrupted speech used in this section and in Section 5.3, and in addition to the noise corrupted speech, multiple speech sources will also be used in the enhancement.

5.5 Speech Enhancement via Separation of Sources from Co-located Microphone Recordings

The speech enhancement algorithms so far presented in this chapter have shown that by taking advantage of the directional characteristics of the AVS array speech enhancement can be achieved at a high quality. In this section the outputs of the AVS array will be used for speech enhancement using the well known FastICA algorithm [161]. Originally the FastICA algorithm was used for source separation applications in anechoic conditions. An extension of the FastICA algorithm, known as the convolutive FastICA [101] was proposed for reverberant recordings, which showed good results in terms of source separation. In general, source separation algorithms have been used for automatic transcription of speech, hands free teleconferencing, speech recognition systems and hearing aids.

For ICA to work efficiently for spatial recordings, there are four essential criteria [99]:

a) Sources originating in different spatial locations should be statistically independent

b) The recordings are made with microphones located at different locations.

c) Each speech signal has a unique temporal structure over short time frames (less than $1s$)

d) The speech signals are quasi-stationary for small time duration ($\sim 10ms$)

Since the AVS array contains co-located microphones the most important criteria in using a BSS algorithm such as the Fast ICA algorithm for sources separation with AVS outputs are the criteria “a” and “b”. The location of microphones is represented in ICA within the mixing matrix; this matrix incorporates information regarding distance and attenuation due to air absorption, and the effects of reverberations on each source to be separated. These characteristics are widely referred to as the acoustic transfer function for each captured signal [100]. In this work, and similar to [162], it is proposed that the mixing matrix in the ICA algorithm should be extended beyond the acoustic transfer functions to include the polar patterns and frequency responses of the microphones used to capture signals. This work investigates the importance of the latter in the ICA mixing matrix for co-located microphones and then considers the consequential impact on enhanced speech quality.

5.5.1 Independent Component Analysis for AVS

The output of an AVS given in (113) consists of four components: an acoustic pressure component and three acoustic particle velocities. In 2D, this can be expressed in vector form as:

$$\mathbf{y}(t) = [o(t) \quad x(t) \quad y(t)] \tag{145}$$

For the gradient microphones, the relationship between the acoustic pressure and the particle velocity is given in (16). The relation between the pressure, particle velocity and the bearing vector has been given in (122 and 123) in Section 4.2.

The traditional ICA model applied to a multichannel speech recording assumes that microphone frequency responses for each channel are the same and that the mixing matrix is a result only of the acoustic transfer function [100]. However, for the AVS, the microphones have directional polar responses. ICA for microphones with directional responses is described in [162]. Following [162] and considering the case of two sources and three microphones (see Figure 69 (a) and (b)), the recorded signals can be modelled using the mixing model:

$$\mathbf{y}(\mathbf{n}) = \sum_{k=0}^{K-1} A_k \, s(n-k) \tag{146}$$

In (146), $\mathbf{y}(\boldsymbol{n})$ represents the digitally sampled microphone signals of (145), $s(n-k) = [s_1(n-k) \quad s_2(n-k)]^T$ represents the vector of source signal samples and $\boldsymbol{A_k}$ represents the convolutive mixing matrices, each of size 3×2. In [162], this model was used to perform ICA on a microphone array containing two closely spaced omni-directional microphones arranged to provide a figure-of-eight polar response and this model is adopted here. In contrast, this work applies ICA to recordings of the acoustic pressure gradient.

In this work, the gradient microphones represented by (122 and 123) are first order and result in figure-of-eight polar patterns as shown in Section 3.3.4. In [163], it was shown that ICA can also be applied to gradient signals, represented by time-differentiated sources signals, and the final outputs are determined by integrating the outputs resulting from separation. The formation of the gradient signals can be modelled as a high frequency boost of the source signals of 6 dB/octave for frequencies above 2 kHz [26] as described in Section 5.3.1. This is similar to applying a pre-emphasis filter, which does not result in a significant change in the perceptual quality of a speech signal. Hence, to avoid approximation errors, the gradient microphone signals of (145) are not time-differentiated prior to applying ICA.

5.5.2 Experiments and Results

Experiments were performed to compare the performance of ICA for speech enhancement using simulated and real recordings from various types of microphone arrays. The experimental setup and the database described in Section 5.2.1 is used in this section.

Anechoic recordings were processed using FastICA [100] while reverberant recordings were processed using a convolutive FastICA algorithm [101]. The resulting separated speech signals were analyzed using the ITU-PESQ software [115] (following low pass filtering and down-sampling to 16 kHz) as described in Section 5.2.2. In addition to the ITU-PESQ software, listening test were carried out according to the setup of Sections 5.2.2. and 5.3.7.

5.5.3 Simulation Experiments

This section examines the role played by the microphone characteristics on the quality of the output produced by ICA using simulated recordings. Simulated anechoic recordings were created using Roomsim [159] and the test database of Section 5.2, with no attenuation due to air absorption and source-to-microphone distances set to 1 m. In all simulations the SNR for the source and interferer were set at 0 dB, corresponding to the worst case scenario.

Three types of co-located microphone arrays were examined. The first array consists of two omni-directional microphones, each with flat frequency responses similar to the Knowles 3132 omnidirectional microphone (refer Figure 20). The second array consists of two omni-directional microphones, one with a flat frequency response (similar to one described above) and one with a frequency response having a 6 dB/octave rise above 2 kHz (matching that of a real gradient microphone [26], Knowles

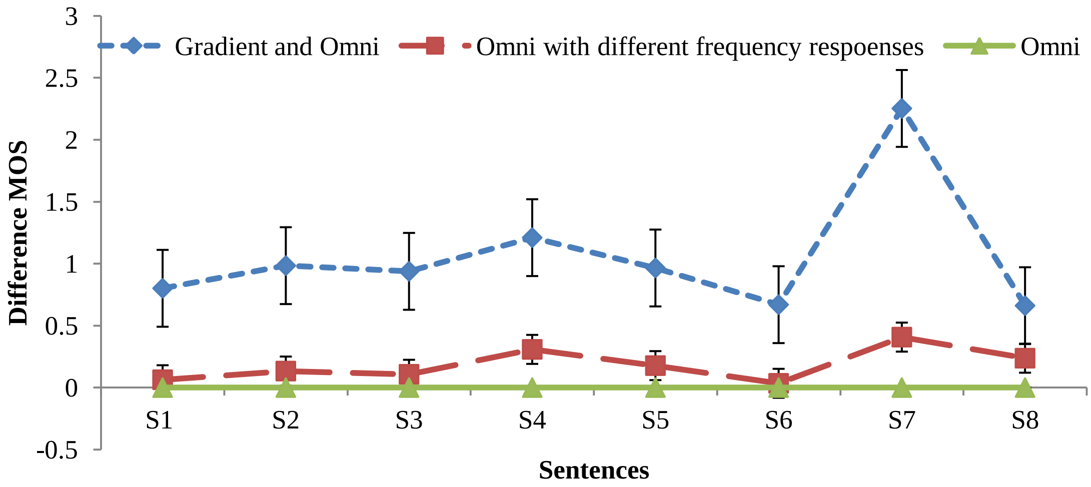


Figure 82: Simulation Results for Omni and Gradient Sensors (Error bars indicate 95% confidence intervals)

NR 3158, refer to Figure 21). The third array consists of one omni-directional microphone with a flat response as the one described above and one gradient microphone having the same frequency response as the second microphone of array two but with the addition of a figure-of-eight polar response (matching that of a real gradient microphone [26], Knowles NR 3158, refer to Figure 21).

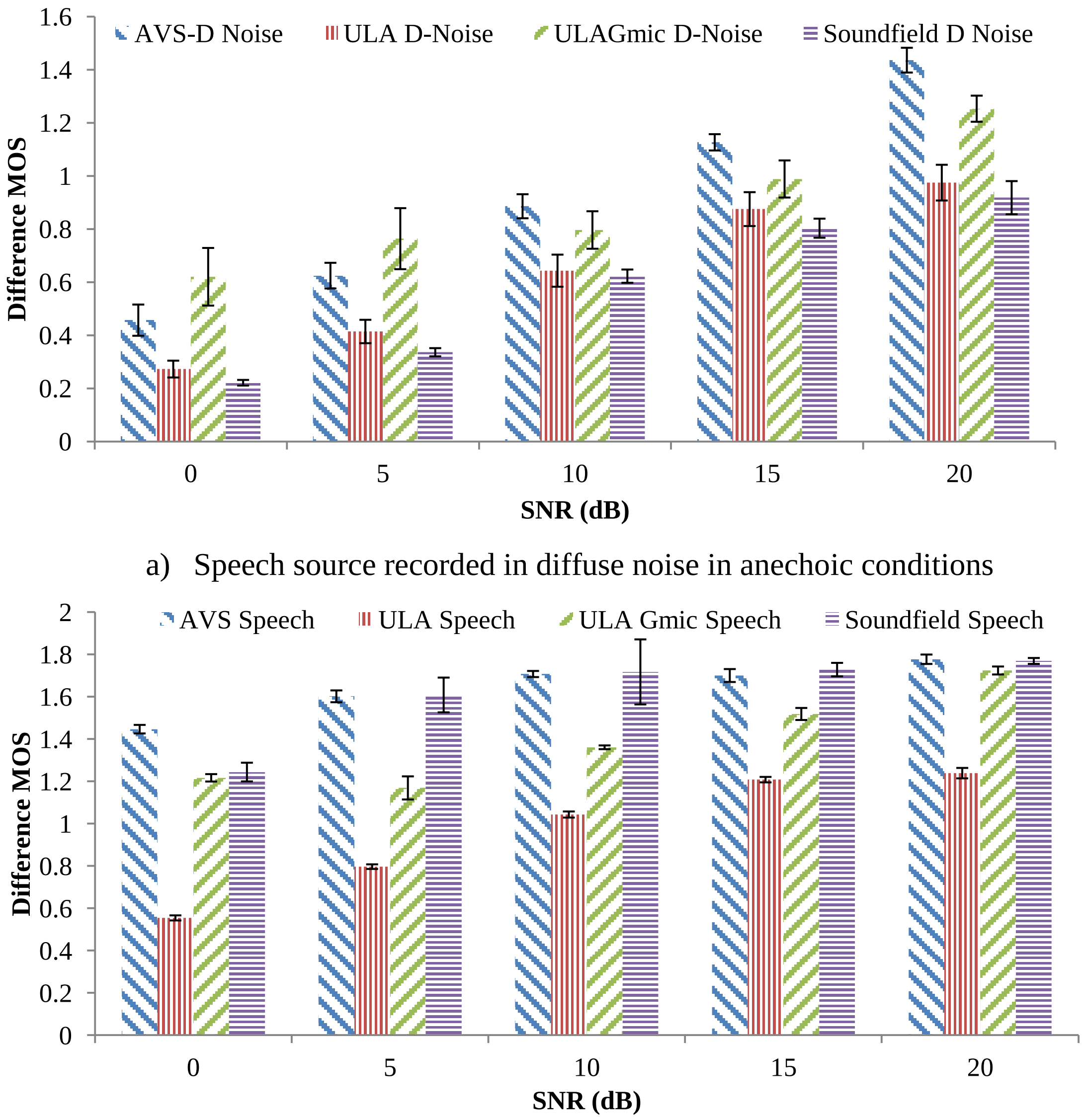
The results obtained from the simulations are shown in Figure 82. There is no improvement in the MOS when using co-located omni-directional microphones with identical frequency responses. For the second array, there is an improvement in the MOS of 0.18. This shows that there is a small contribution to the ICA mixing matrix by the frequency response of the microphone. For the third array which is an AVS simulated with only omni and X sensor, the results of Figure 82 show that there is a significant improvement in MOS of 1.24. These results indicate that the main factor in the performance of ICA for speech enhancement from an AVS is the polar responses of the microphones.

5.5.4 Experiments with Real Recordings

The microphone arrays used for the experiment were:

a) Acoustic Vector Sensor,

b) Uniform linear Array with all omni-directional microphones,



a) Speech source recorded in diffuse noise in anechoic conditions

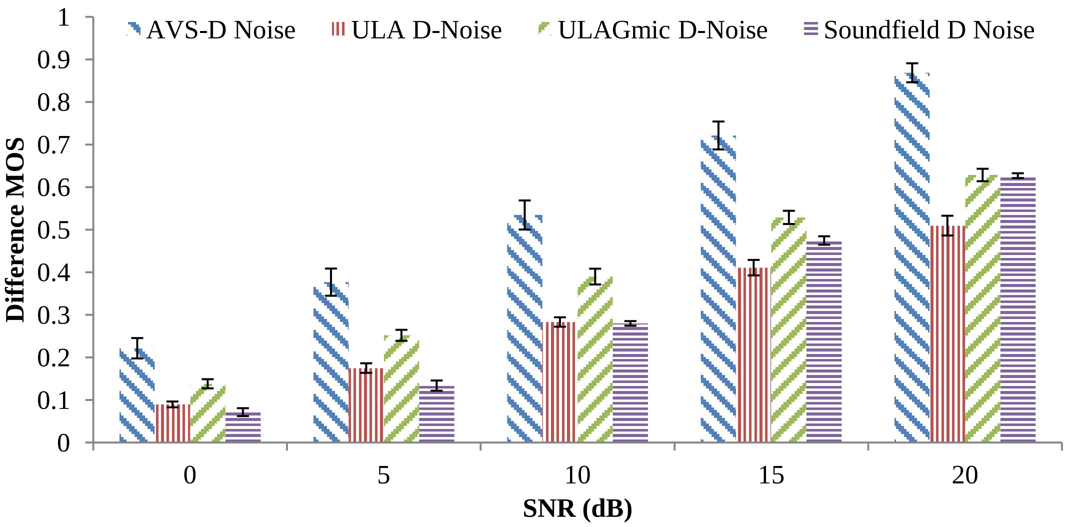
b) Speech source recorded with other speech sources as interferers in anechoic conditions

Figure 83: Results of PESQ MOS for Anechoic room (Error bars indicate 95% confidence intervals)

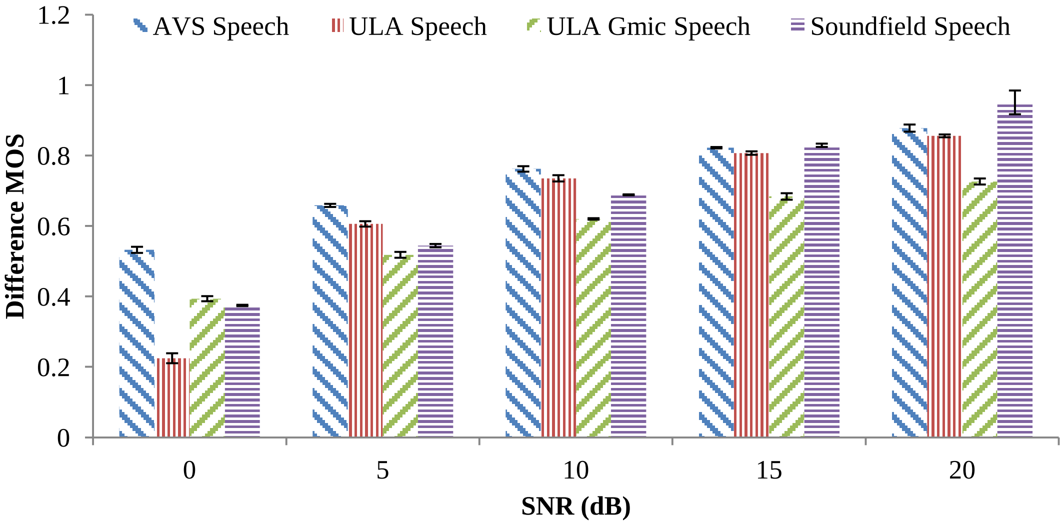
c) Uniform Linear Array with two orthogonally located gradient microphones in x and y planes

d) Soundfield microphone [164] with the polar patterns set to figure of eight.

The ULAs used in these experiments have a length of 300mm with four capsules (either omni or gradient depending on the array) the capsules are spaced 100mm apart, which corresponds to a frequency of 3.4 kHz. Both the AVS and the Soundfield microphones are similar in that they record a 3D soundfield using a co-located array of



a) Speech source recorded in diffuse noise in reverberant conditions



b) Speech source recorded with other speech sources as interferers in reverberant conditions

Figure 84: Results of PESQ MOS for reverberant room (Error bars indicate 95% confidence intervals)

microphones. The key difference between the AVS and Soundfield is the type and arrangement of the capsules.

The results of the experiments for the anechoic conditions are shown in Figure 83 a) and b). For anechoic conditions, with one interferer, the results from processing the AVS recordings with ICA show an average improvement in MOS of 1.65, which is similar to the results obtained from the Soundfield microphone. However, the AVS with 1 speech interferer at an SNR of 0 dB, results in an MOS of approximately 0.2 better than the Soundfield. For diffuse noise, the AVS produces an average improvement in

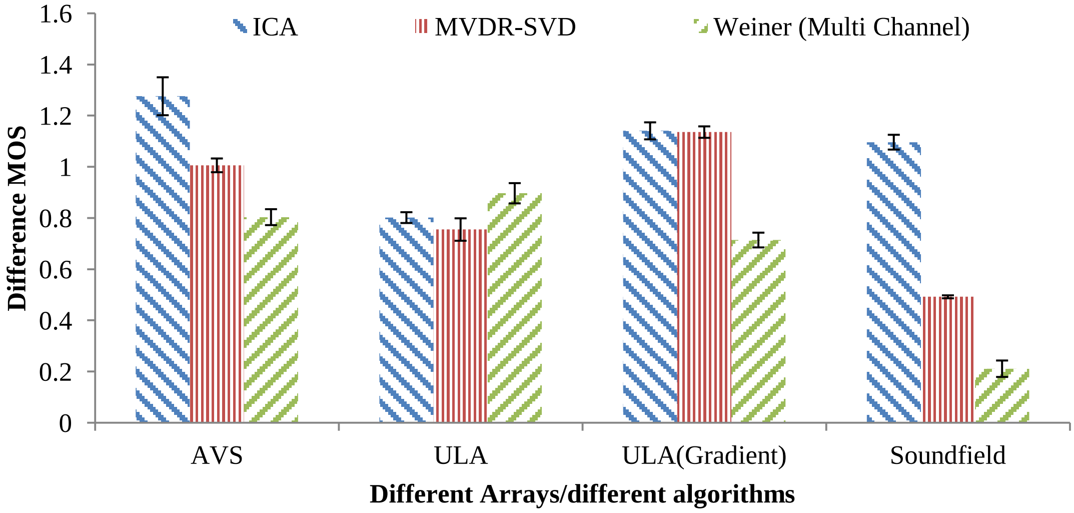


Figure 85: Performance of Different Arrays with different algorithms in anechoic conditions (Error bars indicate 95% confidence intervals)

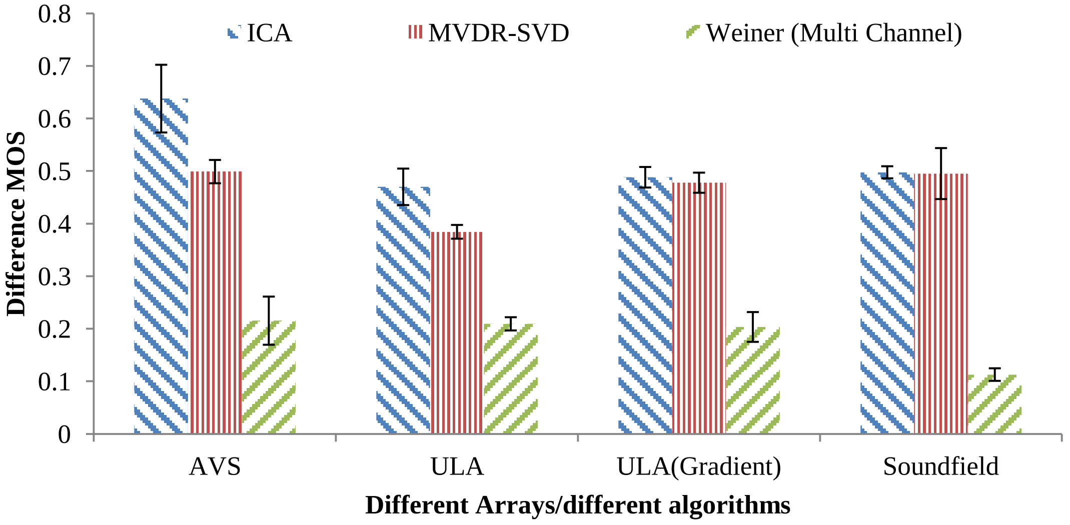


Figure 86: Performance of Different Arrays with different algorithms in reverberant conditions (Error bars indicate 95% confidence intervals)

MOS over all noise scenarios of 0.9, which is similar to the next best performing array (in this case, the ULA with gradient microphones). However, the AVS is significantly better at high SNRs, while decreasing in performance at low SNRs.

The results for the reverberant room are shown in Figure 84 a) and b), where the results are different to those of the anechoic case. For the speech interferer, MOS results for the AVS are on average 0.1 better over all SNR scenarios than the next best performing array, in this case the ULA. For diffuse noise, the AVS again performs better than all other arrays, with an average MOS improvement of 0.14 higher than the next best performing array (again being the ULA). However, for both single interferers and diffuse noise at 0 dB, the AVS performs significantly better (on average 0.4) compared with the ULA, which is the next best performing array.

5.5.5 Comparison of ICA for Speech Enhancement with other Enhancement Algorithms

The results presented in the previous section have shown that by applying the Fast ICA on the outputs of arrays that contain direction microphones an improvement in perceptual quality can be achieved. Here the performance of the ICA for speech enhancement is compared against MVDR-SVD beamformer and Multichannel Weiner filter. The results are presented in Figure 85 and Figure 86 are for anechoic and reverberant conditions for combined noise and speech as interferers. From the results it can be seen that when only gradient microphones are used in the construction of the array, the performance of the enhancement algorithm is better than those arrays with omni-directional microphones and cardioid microphone capsules.

5.5.6 Experiments with Changing Microphone Array Orientation

Having established that the AVS is the best performing microphone array, compared with other scenarios, experiments were conducted to evaluate the impact of array orientation (relative to the source) on the resulting ICA performance. Since the gradient microphones are highly directional the performance of ICA may be due to directing the microphones directly at the source or interferer. The signal to noise ratio is set at 0 (the worst case scenario) and the recordings are made for a single speech interferer. The arrays are rotated in azimuth through 90 degrees at 15 degrees intervals and recordings made for each orientation.

The results in Figure 87 show MOS results for outputs from ICA performed on these recordings in both anechoic and reverberant environments. The results show that there is no effect on the performance of ICA by turning the array in azimuth.

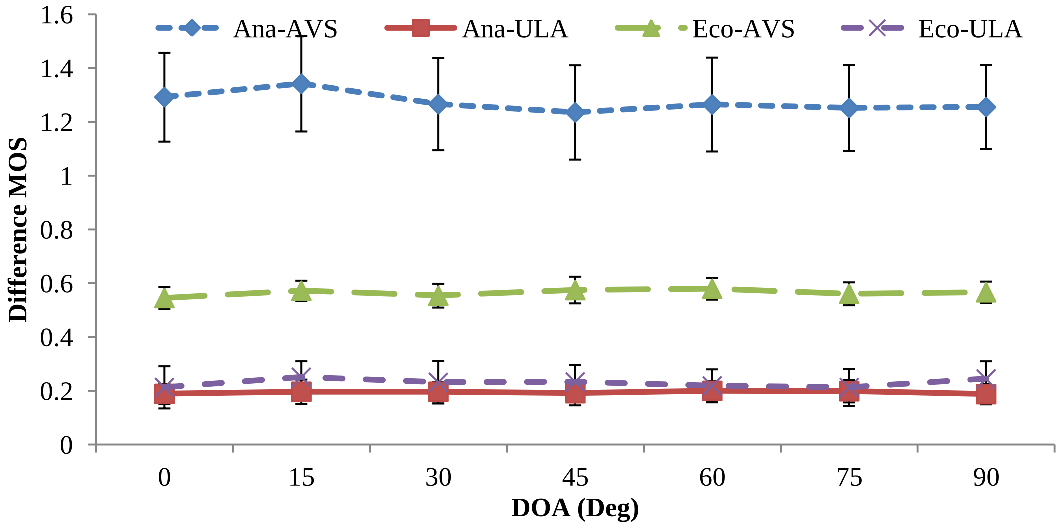


Figure 87: The Difference MOS results for different azimuth angles (Error bars indicate 95% confidence intervals).

5.5.7 Results for Listening Test

The listening tests were divided in two parts. The listening tests for noise corrupted speech which is presented here and results for mixed speech sources which will be presented in Section 5.6.8. The listening test carried out here was conducted as described in Section 5.2.2 and 5.3.7. The result for the ICA for listening quality is compared with the listening quality of the MVDR-SVD beamformer and the perceptual filter presented in Section 5.4.

The results in Figure 88 show that the listeners scored all the algorithms tested equally, accept for the unprocessed results. The results show that listeners ranked the output from the algorithms fair (~3.0), and the unprocessed files as bad (~1.6). Hence, it can be concluded that the performance of the ICA for enhancement of speech sources corrupted by noise performs equally to that of the MVDR-SVD beamformer, GJ beamformer and the perceptual filter proposed in Section 5.4.

5.5.8 Discussion

The work presented in this section has shown that there is a significant impact on the performance of ICA applied to co-located microphone arrays when the microphones of different polar responses are used. This agrees with previous work

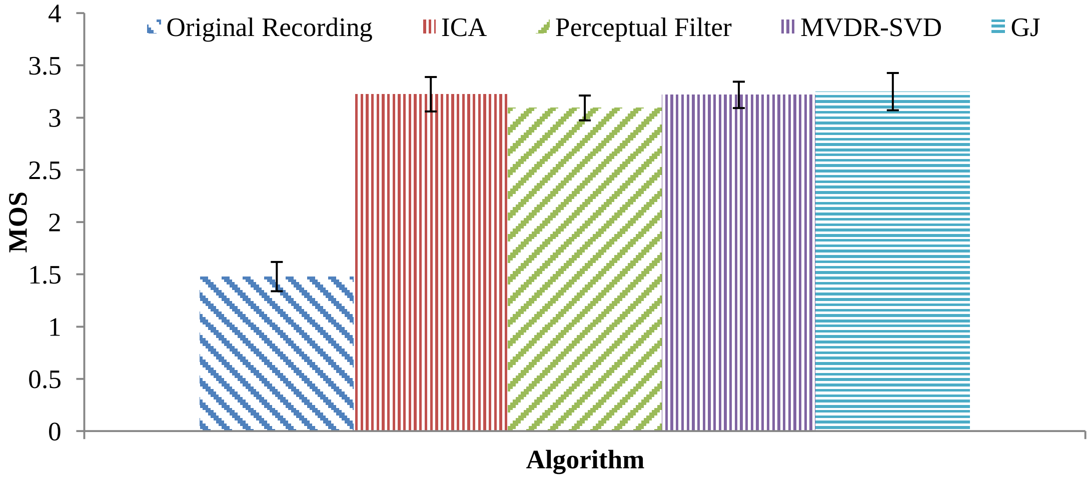
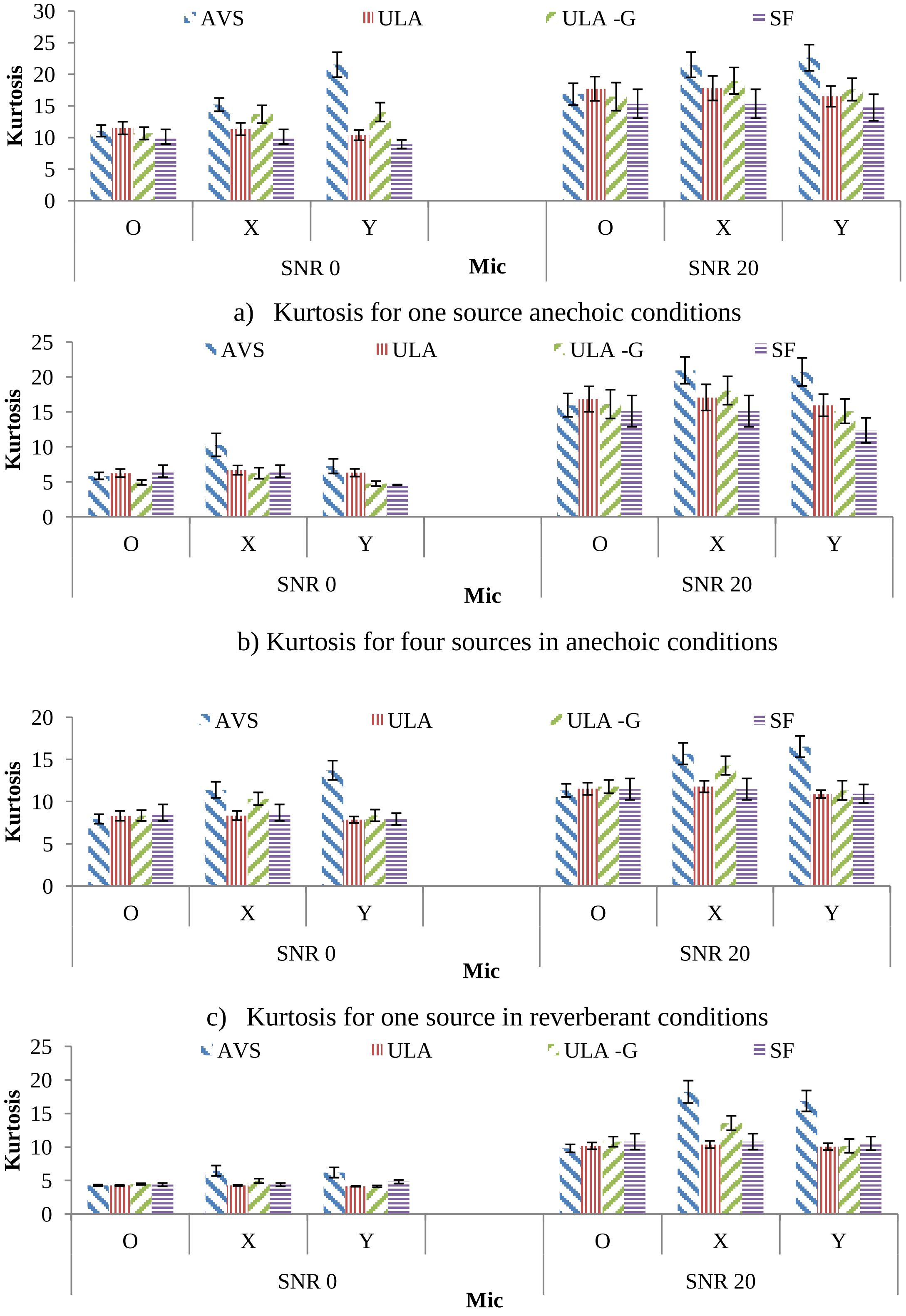


Figure 88 : Results for listening test of ICA compared with other filters (Error bars indicate 95% confidence intervals).

investigating ICA for closely spaced microphone arrays with directional responses [163]. It is suggested that using directional microphone recordings results in increased statistical independence between the recorded signals and, in turn this results in improved separation performance using ICA. Using the database of recordings described in Section 5.2, the kurtosis and mutual information of each of the microphone recordings was measured, with the results shown in Figure 89 and Figure 90.

The results show that when the microphones have directional polar responses, the recorded signals from different channels will have significantly different kurtosis and mutual information values will be significantly less between channels. When results from Figure 89 and Figure 90 are compared with the performance results of ICA in Figure 83 and Figure 84, it is seen that where there is a large variation in the kurtosis and values of mutual information is small between the channels, the performance of ICA as measured by PESQ is improved. This indicates that directional microphones result in signals that are more suitable for separation via ICA, compared with arrays of non-directional microphones.



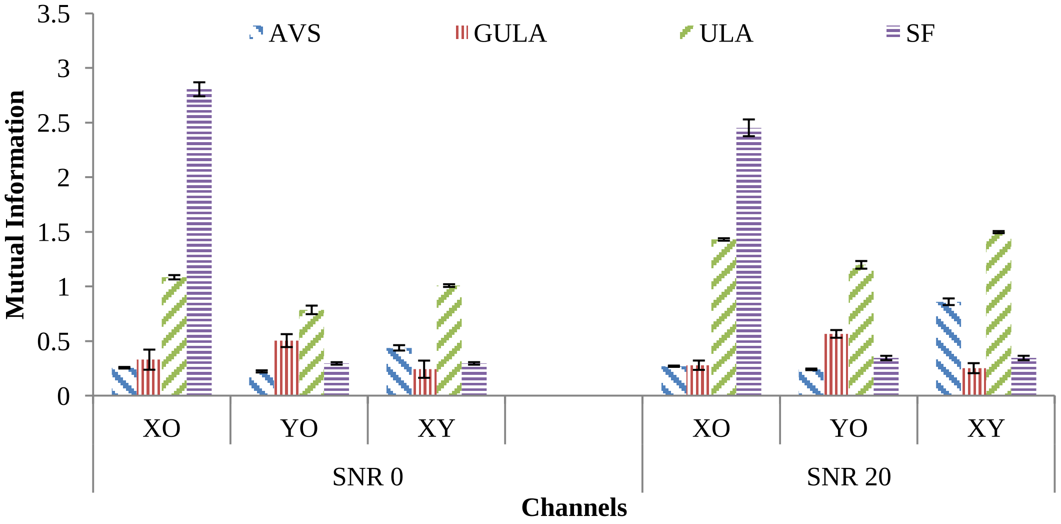
a) Kurtosis for one source anechoic conditions

b) Kurtosis for four sources in anechoic conditions

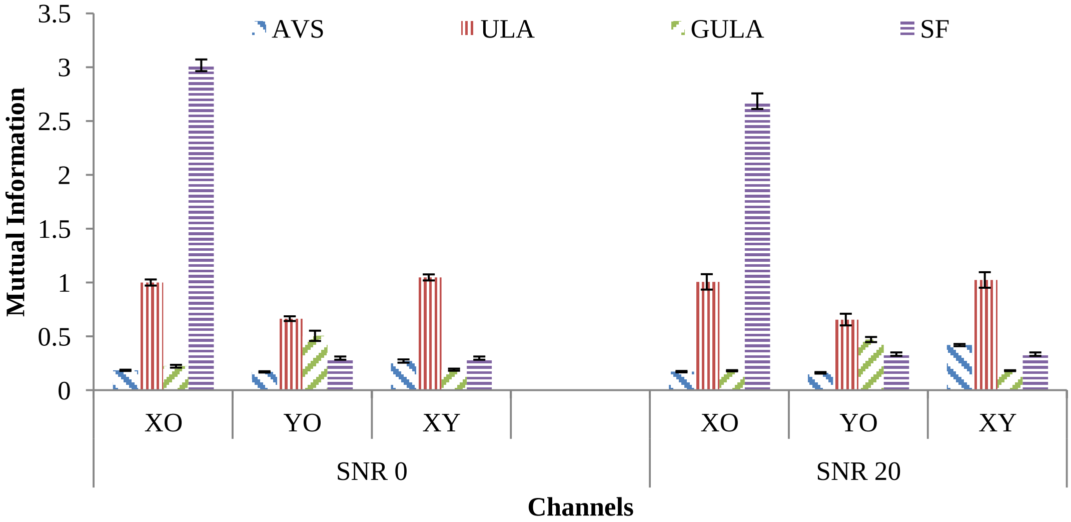
c) Kurtosis for one source in reverberant conditions

d) Kurtosis for four source in reverberant conditions

Figure 89: Kurtosis for Channels of the AVS array (Error bars indicate 95% confidence intervals)



a) Mutual Information for channels of different arrays in anechoic conditions



b) Mutual Information for channels of different arrays in reverberant conditions

Figure 90: Mutual Information for Channels of the AVS array (Error bars indicate 95% confidence intervals)

5.5.9 Summary

This section has investigated speech enhancement using source separation techniques applied to an AVS. Source separation is based on a convolutive ICA model applied to co-located microphones that both record omni-directional with directional gradient signals. Perceptual quality results (measured using PESQ) show a significant improvement in speech enhancement using ICA applied to an AVS compared to ICA applied to a traditional linear microphone array, in both anechoic and reverberant environments. In addition a comparison was made between ICA, MVDR and multichannel Weiner filter for different arrays. This comparison showed that the performance of the AVS is significantly better than other arrays tested. Furthermore, the results showed that when pressure gradient sensors are used, the performance is better for all algorithms tested in the experiment. Results also show that a key factor in the performance improvement is the use of directional polar responses, which lead to recorded signals that are more statistically suited to ICA.

The result presented in this section has shown that a co-located microphone array such as the AVS can be used with a source separation algorithm such as the Fast ICA. In the next section, a method for sources separation based on the DOA estimates from Chapter 4 will be presented.

5.6 Separation of Speech Sources Using an AVS: Beyond ICA

There are numerous proposals for solving the problem of BSS [99] applied to speech signals. Many existing approaches to BSS are not suited to real-time multimedia applications such live audio ‘browsing’ of hands-free meeting recordings or remote teleconferencing, where the objective is to allow selective enhancement of a desired speaker in a multiple participant scenario. Further, these applications typically operate in reverberant environments for which BSS solutions focusing on convolutive mixing of sources is required. Here, a new technique for BSS is proposed that can operate on a single 20 ms frame of speech and is thus well suited to such multimedia applications. This technique relies on the DOA estimates from the AVS. The AVS provides highly accurate estimation of sound source directions as shown in Chapter 4 and successfully used in the enhancement of single speech recordings in realistic reverberant environments in Sections 5.3 to 5.5. Here, a BSS algorithm for separating mixed speech recordings based on exploiting these directional characteristics of an AVS array is proposed.

In the previous section, ICA, one of the most commonly used BSS algorithms for multiple microphone recordings [100], was used for speech enhancement of AVS recordings in anechoic and reverberant environments. A more recent BSS algorithm that has shown good performance in reverberant environments is the Degenerate Un-mixing Estimation Technique (DUET) BSS algorithm [165]. The idea behind DUET BSS is that it assumes that sources are W-disjoint orthogonal in the time-frequency domain and

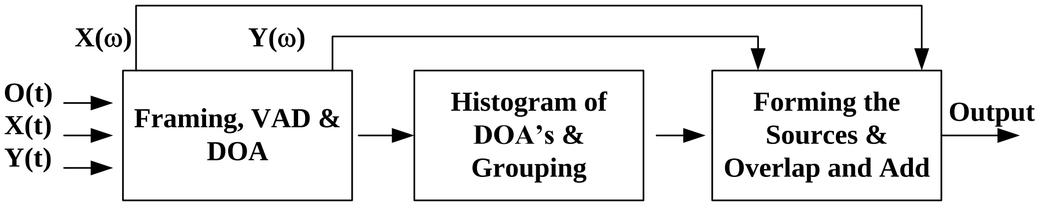

X(ω)
Y(ω)
O(t)
X(t)
Y(t)
Framing, VAD & DOA
Histogram of DOA's & Grouping
Forming the Sources & Overlap and Add
Output

Figure 91: Block diagram of the proposed method

by partitioning the time frequency representations of the mixtures, the sources will be separated. Furthermore in [166] a method based around an expectation maximization algorithm and frequency bin wise clustering is presented, and in [167] an extension to the DUET algorithm is presented where time frequency components of the signals captured from randomly arranged microphones are clustered based on TDOA estimates and time frequency masking is used for source separation.

The approach proposed in this section is similar to this latter approach in that it aims to separate individual time frequency sources based on estimates of their location relative to a microphone array. However, unlike many of the previous TDOA-based approaches that are based on spaced microphone arrays (e.g. [48, 168]), the AVS, being a co-incident microphone array, requires an alternative method for estimating the source directions. Hence, the method chosen here is similar to the BSS technique recently proposed for the Soundfield microphone [137].

Here, the DOA estimation approach of Section 4.5 is used, resulting in estimations of source directions on a time-frequency basis. Similar to other approaches [137, 166-168], the proposed technique then achieves source separation by grouping time-frequency components with similar DOA estimates into individual sources. Compared to [137], a key difference in this work is that compared to the Soundfield microphone, the AVS consists of different types of microphones, in this case gradient sensors rather than pressure sensors. In previous sections it has been shown that processing of the gradient signal recordings provides clear advantages for both speech enhancement and sound source localisation in reverberant environments using only short time frames (10 ms) of speech signals as shown in Chapter 4. Further, this work also introduces a technique similar to dithering [169] to reduce musical distortion in the separated speech signals, a common problem with time-frequency approaches to BSS.

5.6.1 Source Separation Based on the DOA Estimates

The DOA estimation approach of Section 4.5 results in a direction estimate for each time-frequency component. This section describes the method for separating mixed speech sources by combining time-frequency components with similar direction estimates. The block diagram of Figure 60 can be extended as shown in Figure 91 to include the source separation block. The first block of Figure 91 remains the same as that described in Section 4.5. The changes to the DOA estimation algorithm are made in the block where the histograms and groupings are formed.

5.6.2 Histogram of the DOAs in the FFT Bins and Grouping for Source Separation

The method of clustering DOA estimates from FFT bins was described in Section 4.5, where only the DOAs where clustered. Here, once the correct histograms of the DOAs are formed, the sets of corresponding frequency components denoted K_i are also clustered with the DOAs. The result of this processing is the formation of binary time-frequency masks for each source, as described by:

$$M_i(t,k) = \begin{cases} 1 & k \in K_i \\ 0 & otherwise \end{cases} \quad i = 1, \ldots, M, k = 0,1, \ldots, N \tag{147}$$

where M is the number of unique sources and where for Figure 61, $M = 3$. These masks are then used to form the separated sources as described in the next section. The result of the clustering is illustrated in Figure 61 (b), which shows the total number of frequency coefficients identified for each of the three sources.

5.6.3 Forming the Sources in each Look Direction

The frequency components that are used in the reconstruction are from the x and y gradient components of the AVS output. It was found that when the omni-directional recording is used, the performance of the separation is not as good as when only the x and y components are used (a similar result was found in previous sections in this chapter). The separated sources are obtained using the mask of (147) applied to the time-frequency components from the gradient recordings $X(t,k)$ and $Y(t,k)$ using:

$$S_x^i(t,k), = M_i(t,k)X(t,k) \tag{148}$$

$$S_y^i(t,k), = M_i(t,k)Y(t,k) \tag{149}$$

The result of this process is the formation of separate source signals as given by $S_x^i(t,k), S_y^i(t,k)$, which represents the time-frequency components taken from $X(t,k)$ and $Y(t,k)$ that represent the i^{th} separated source.

Since (148) and (149) applies a binary mask, some components from the original frame of speech may be missing (e.g. if two sources had simultaneous time-frequency components). This can result in the presence of musical distortion in the separated sources, especially in those directions where there are weaker sources. To minimize this musical distortion, a proportion of the time-frequency components from the other sources are added back to the time-frequency regions that were zeroed for the current source in (148) and (149). This can be expressed as:

$$\tilde{S}_x^i(t,k) = S_x^i(t,k) + Dis \sum\nolimits_{j \neq i} S_x^j(t,k) \tag{150}$$

$$\tilde{S}_y^i(t,k) = S_y^i(t,k) + Dis \sum\nolimits_{j \neq i} S_y^j(t,k) \tag{151}$$

where Dis is the scaling factor for the missing frequency components, which in this work, was varied between 0.175, 0.2, 0.25 or 0.3 and results compared in section 5.6.4. The final time-domain separated speech signals are obtained using overlap add reconstruction and adding together the *x* and *y* components of each source as described in (137) with $\beta = 0$.

5.6.4 Experimental Setup

The database used in this work is different to that described before in Section 5.2, where the database is composed of speech sentences which are 6s long taken from the TIMIT database[170]. The sentences included 12 male and 12 female speakers. Recordings were made at a sampling rate of 48 kHz and then down-sampled to 16 kHz before being processed. In total, 36 recordings were made for the three speaker case and 18 recordings for the two speaker case. The combinations of the speakers were arranged such that samples of all male, all female, and male and female could be obtained. The case of all male or all female is considered a harder problem than samples of male and female speakers, when the all male or female samples are used, frequency content of the two sources are in the same regions hence there are more errors in the DOA estimation. Three self powered loud speakers (Genelec 8020A) were arranged in front of an AVS as shown in Figure 92. The outputs of the loud speakers were set to 90 dBA. The recordings were made in a normal meeting room with a RT_{60} of 30ms.

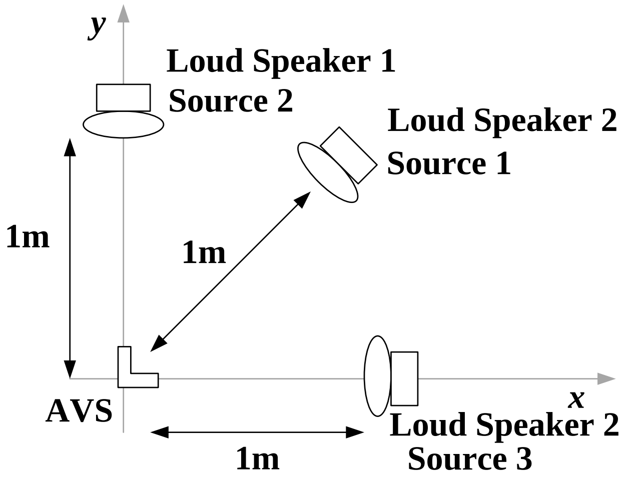


Figure 92: The experimental setup

In most BSS algorithms, the performance evaluation of the algorithms is based on SIR and SDR as described in (102 and 103). These measures provide a good estimate of how much of the target source as been separated from the mixed sources. Unfortunately, these methods fail to give an indication of the perceptual quality of the separated sources. Hence, in this work, the source separation performance was measured on the basis of subjective and objective perceptual evaluation of the separated sources as well as the traditional measures of SDR and SIR.

The SDR and the SIR of the unprocessed, output of the ICA and the output of the proposed algorithm was calculated using the BSS evaluation tool kit proposed in [117]. The SDR and SIR for each channel were calculated against the original recordings without any interferers. The improvements in the SDR and SIR were calculated from the difference in SDR and SIR between the results for the unprocessed and the processed outputs.

The PESQ MOS test were carried out as described in Section 5.2.2, in addition to the PESQ, a MOS listening test of the separated signals was carried out according to [114]. The listening tests include twenty listeners all native English speakers (ten male and ten female). The listening tests contained four sets of the separated speech files, with each set containing the separated files from ICA, the original without interferers, the corrupted speech files and the separated files from the new algorithm. In total there were 48 files, 6 s long played in a randomized order to listeners.

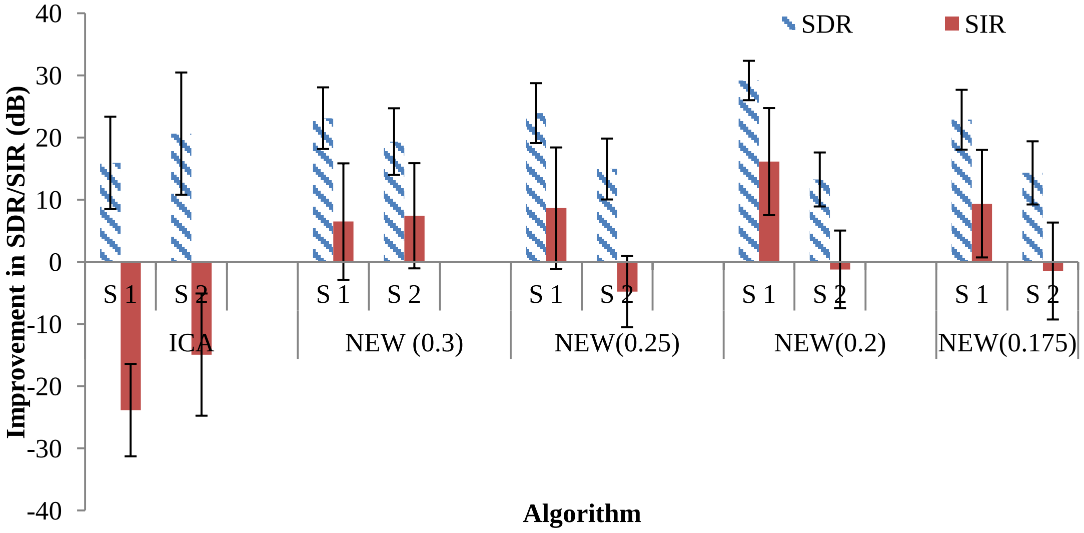


Figure 93: Improvement in SDR and SIR over the unprocessed recordings for two sources (Error bars indicate 95% confidence intervals)

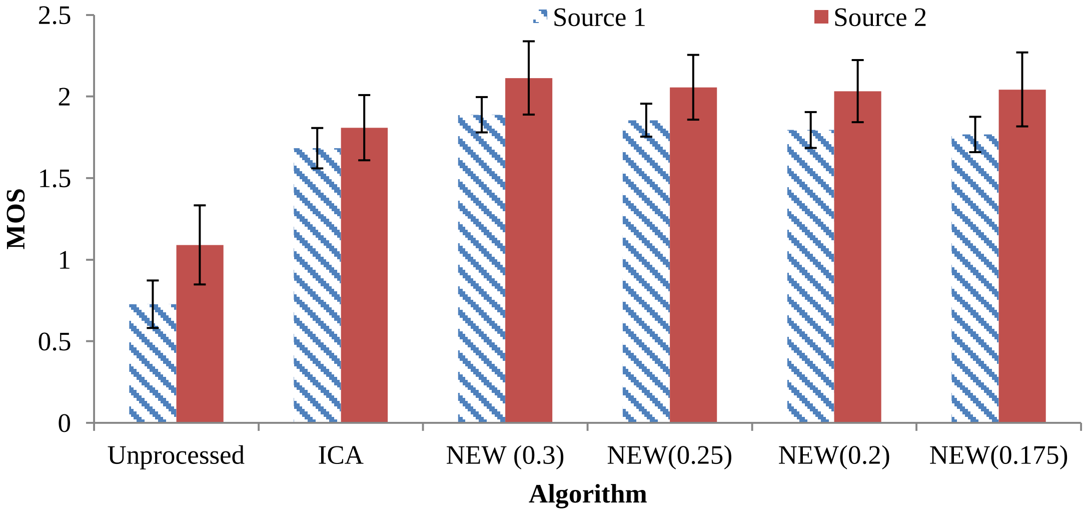


Figure 94: MOS for two speakers (Error bars indicate 95% confidence intervals)

5.6.5 Experiments with Real Recordings

The results presented are for sources located at 0, 45 and 90 degrees; since the new algorithm produced outputs for 22 look directions (8 degree resolution) only the outputs that correspond to the actual source directions are used in the evaluation of the performance of the proposed algorithm. In the case of the two sources, only the outputs corresponding to 0 and 90 degrees are presented. The results presented are for the source separating algorithm described in this work (at differing levels of restoration of the missing frequency components) against the unprocessed and the ICA algorithm as a benchmark.

5.6.6 Results for Two Sources

The results for the SDR and SIR shown in Figure 93 shows that there is an average improvement of 24.8 dB and 15.5 dB over the unprocessed recordings in terms of SDR and 10.15 db average improvement in one source over the unprocessed recordings. In the case of the second source, when the ratio of adding the missing frequency is at 0.3, there is an improvement of 7.4 dB over the unprocessed recordings, but when the ratio is reduced the SIR for that channel goes down. Overall, the SDR for all scenarios increased over the unprocessed recordings. For ICA, there is no improvement in terms of SIR but there is an improvement of 16 and 20.5 dB improvement over the unprocessed recordings in terms of the SDR.

The results in Figure 94 show that there is an average improvement of 1 and 0.7 MOS for ICA compared to that of the unprocessed recordings. In comparison with ICA, the new algorithm produces an improvement of 1 MOS for each of the channels over the unprocessed recordings. The results show that the new algorithm produces better improvements to both sources where as in ICA one channel (the dominant channel) is improved more than that of the other channel. From the results, it can be seen that the amount of restoring the missing frequency components has very little effect on the performance according to PESQ score.

5.6.7 Results for Three Sources

The results presented in Figure 95 are for the improvement in SDR and SIR for ICA and the proposed technique over the unprocessed signals. The results show there is an average improvement of 4.3, 0.2 and 11.6 dB over the unprocessed recordings for SDR from the proposed algorithm. For ICA in terms of SDR, there is no improvement and in fact the processed files from ICA showed the SDR got worse by 15.4, 17.5 and 10dB for the three sources over the unprocessed recordings. In the case of SIR, the average improvement over the unprocessed recordings for the proposed algorithm is 14.8, 8.2 and 22.2 dB for the three sources. ICA shows an improvement over the unprocessed recordings of 9.5, 8.5 and 16.5 dB improvements for each source. From the

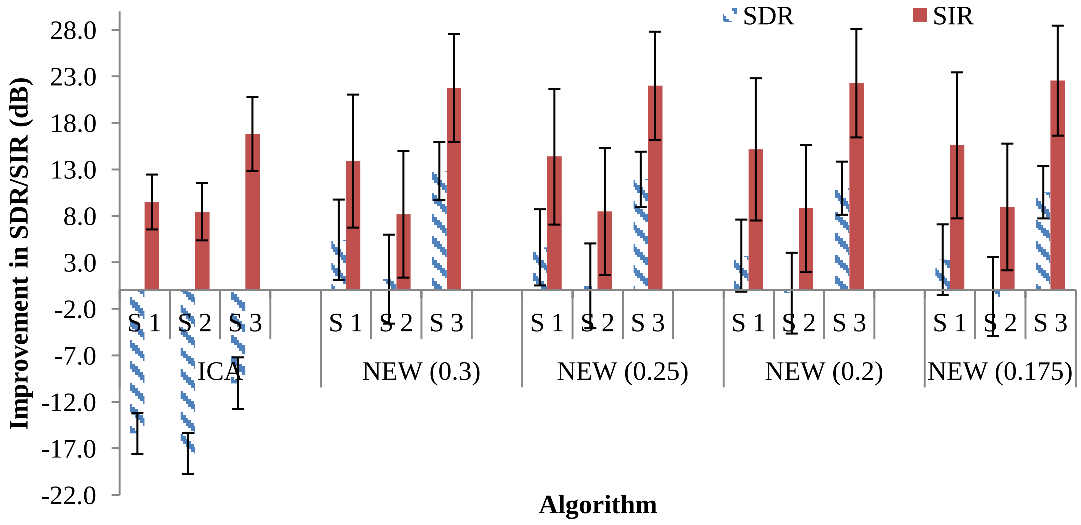


Figure 95: Improvement in SDR and SIR over the unprocessed recordings for three sources (Error bars indicate 95% confidence intervals)

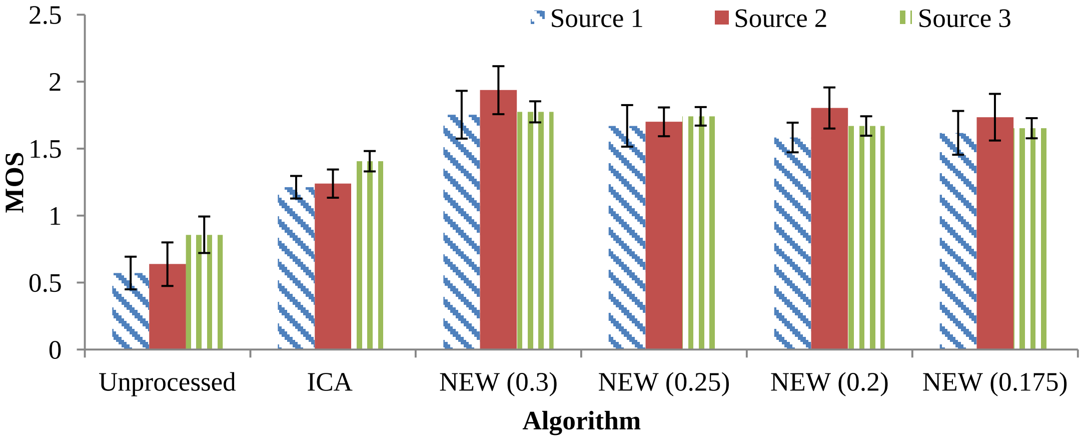


Figure 96: MOS for three sources (Error bars indicate 95% confidence intervals)

results it is seen that when the level of restoration of the missing frequency components is reduced, there are better SIR improvements but the SDR becomes worse. This is expected as by reducing the contributions from the frequency components from other directions will increase the amount of musical distortion.

The results presented in Figure 96 are for the case of three consecutive speakers. The improvement in terms of MOS compared to the two speaker case is much more significant. Overall the new algorithm showed improvement in MOS of 1.1, 1.2 and 0.9 over the unprocessed channels and ICA approach produced an improvement of 0.6, 0.6 and 0.5 over the unprocessed channels. The improvement in performance between ICA and the new algorithm is on average 0.6 for the three sources case. This is more significant than in the case where there are only two speakers. The results show that when the number of sources is increased, the performance of ICA is reduced whereas the performance of the new technique remains constant in terms of PESQ and the performance in-terms of SDR are SIR is more significant than the case of the two sources. In addition, the results show that the proposed algorithm is able to produce improvements for all three separated sources.

5.6.8 Listening Tests for the Proposed Filter Outputs

The results of the listening tests using real listeners are presented in Figure 97. The results from listening tests shows that the new algorithm produces an increase in MOS of 1.3, 1.0 and 1.6 over the unprocessed recording and for ICA the MOS is 0.4.0.4, 0.9 over the unprocessed recordings. Although the overall pattern of the results is similar to that of the SIR, SDR and PESQ results, the actual MOS scores are higher for the listening tests compared to that of the PESQ. In the case of the new technique, the listeners ranked the improvement from poor to fair for two channels and poor to good for one channel. In contrast, the results from PESQ show that improvement for the new technique is from bad to poor. This highlights the need for real listening tests to evaluate the performance of source separation and enhancement techniques.

5.6.9 Summary

This section has described a new technique that can be used for source separation based on the use of an AVS to identify unique spatial locations of individual speech sources in the time-frequency domain. Objective and subjective testing verifies that the approach achieves high quality source separation using just a single 20 ms frame of speech. In the next section, these algorithms will be combined with the LP perceptual filter presented in Section 5.4.

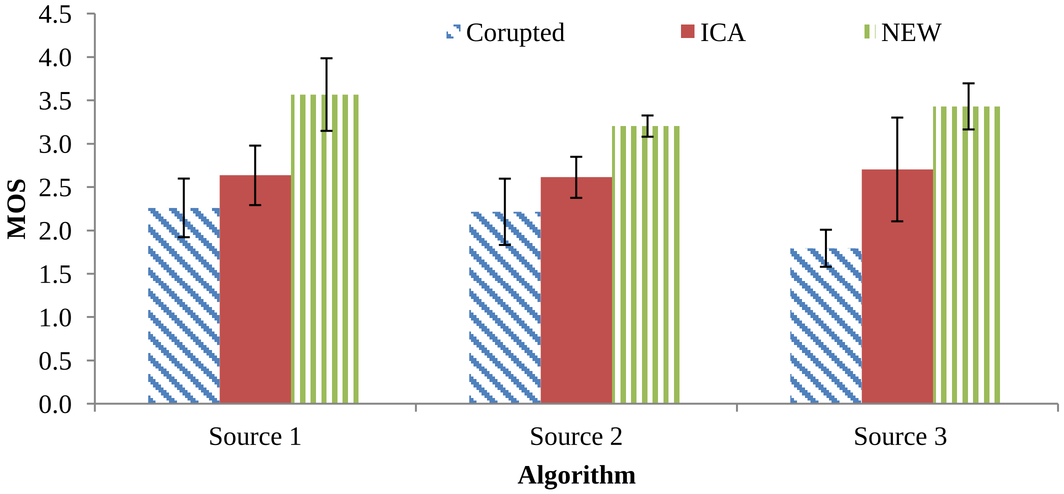


Figure 97: Results of MOS listening test for three sources (Error bars indicate 95% confidence intervals)

5.7 Methods of Obtaining Accurate LP Spectra for Perceptual Filtering

In Section 5.4 the LP spectra based speech enhancement filters for the AVS resulted in a significant improvement in terms of PESQ scores and MOS listening tests. The fundamental component of the filter proposed in Section 5.4 is based on the LP spectra of the speech signal. In noisy and reverberant environments, degradation in signal quality leads to inaccurate estimation of the LP spectrum [171, 172], which in turn will reduce the performance of the speech enhancement algorithm. LP spectra obtained from microphone arrays have been shown to be more accurate for reverberant speech and in [173] and in [174] recordings from a microphone array were used to increase the SNR of the signals in order to get a better coding performance. Furthermore, a microphone array offers superior performance over a single microphone in reducing signal distortion and speech intelligibility degradation resulting from noise removal [28]. In this section different multichannel processing techniques will be applied to the outputs of an AVS to obtain an accurate estimate of the LP spectra of the speech signal corrupted by noise.

There are several ways of processing multichannel signals to form LP spectra. These include beamforming the signals to form a single enhanced channel, obtaining the average autocorrelation of all signals, which is then used to form the LP spectra [29], use of BSS techniques to enhance the noise corrupted speech signal and finally using the MVAR modelling for multichannel LP analysis [94]. In this section it will be shown that by using the AVS gradient channels with the methods described above an improved estimate of the LP spectra can be obtained for use in the speech enhancement algorithm of Section 5.4.

5.7.1 Methods for Enhancement of LP Spectra of Noise Corrupted Speech

The effect of noise on the LP spectra of speech was shown in Figure 74 (a) Section 5.4.3. It was shown in Section 5.4 that by combining the outputs of the gradient microphones, a more accurate estimate of the LP spectra can be obtained. Here, different algorithms used in combining the gradient sensors to one channel which can then be used in calculating the LP spectra are proposed.

A combined LP spectrum from multiple channels can be obtained by four methods in general.

a) Beamforming the AVS channels to obtain a single outputs and performing LP analysis.
b) LP spectrum estimation using the averaged autocorrelation matrix of all the AVS channels.
c) Using BSS techniques such as ICA to enhance the source signal.
d) Using the MVAR approach.

5.7.2 Beamforming the AVS Channels

In Section 5.3, four methods for beamforming were presented. In this section these beamforming methods described in Section 5.3 will be applied to the output of the AVS array before the calculation of the LP spectra in the filtering process. Since the results of applying the beamformer of the output of the AVS shows improvements in perceptual quality, using those outputs will give a closer estimate of the LP spectra.

5.7.3 LP Spectrum from Averaged Autocorrelation Matrix of all the Channels

The output channels of the AVS array will be used to get an LP spectrum of the combined channels using the averaged autocorrelation matrix method briefly described in Section 2.9. Here, a detailed explanation of the averaged autocorrelation for obtaining the LP coefficients will be presented.

The observations of the channel output from any array in terms of linear prediction can be expressed as[29]:

$$z_m(n) = \sum_{i=1}^{p} a_{m,i} z_m(n-i) + e_m(n) \quad m = 1,2 \cdots M \tag{152}$$

where $z_m(n)$ is the m^{th} channel output, $e_m(n)$ is the LPC residual obtained from m^{th} channel and a_m is expressed as[29]:

$$\boldsymbol{a}_m = \boldsymbol{R}_{zz,m}^{-1} \boldsymbol{r}_{zz,m} \tag{153}$$

where

$$\boldsymbol{a}_m = \left[a_{m,1}\ a_{m,2} \cdots a_{m,1}\right] \tag{154}$$

and $\boldsymbol{R}$ and r are autocorrelation matrix and first column of the autocorrelation matrix defined as [29]:

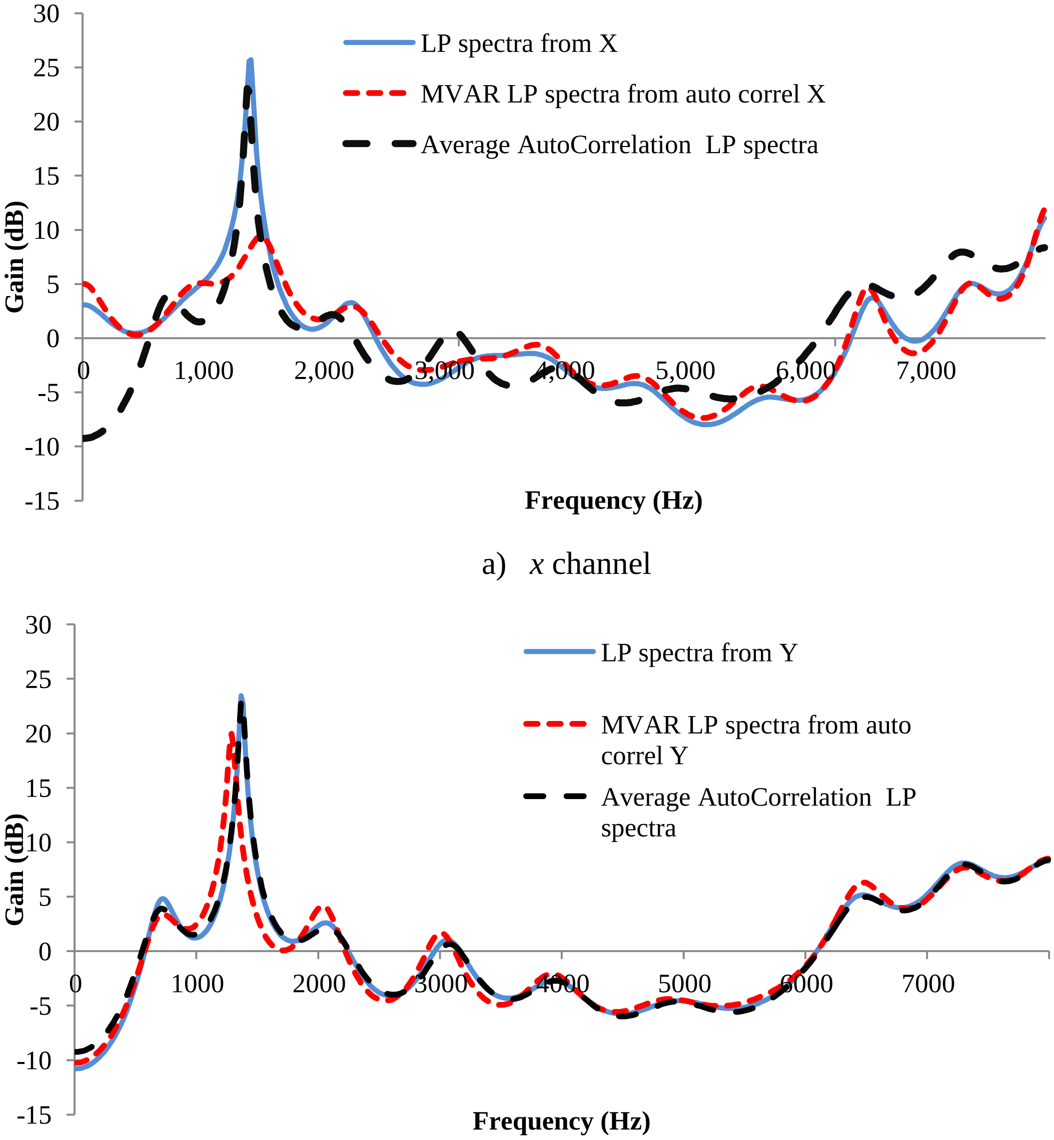
$$\boldsymbol{R} = \begin{bmatrix} r_{zz,0} & r_{zz,1} & \cdots & r_{zz,p-1} \\ r_{zz,1} & r_{zz,0} & \cdots & r_{zz,p-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{zz,p-1} & r_{zz,p-2} & \cdots & r_{zz,0} \end{bmatrix} \tag{155}$$

The averaged autocorrelation function for the M channels can be described as[29]:

$$\boldsymbol{R}_{zz} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{R}_{zz,m} \tag{156}$$

In Section 5.4, the LP spectral for the postfilter is obtained from the output of the summing beamformer described in (137). In the averaged autocorrelation method, the x and y components of the AVS output are used in the formation of the averaged autocorrelation function which is then used in the perceptual filter described in Section 5.4 to form the LP spectra for filtering.

Unlike the implementation of the original algorithm of Section 5.4 there is no beamforming for LP spectra estimation, which was then used in the filtering. Here, with the autocorrelation method for obtaining the LP spectra the summing beamformer is used to combine the channels which are then filtered with the LP spectra, from the averaged autocorrelation function.



a) *x* channel

b) *y* channel

Figure 98: LP spectra from autocorrelation of *x* and *y* channels, LP spectra from LP coefficients from autocorrelation in MVAR and LP spectra from averaged autocorrelation of both channels

The plot of the LP spectra from the average autocorrelation method is shown in Figure 98. It can be seen that the LP spectra based on the average autocorrelation function is a closer representation of the LP spectra of the clean signals in both the *x* and *y* channels. Furthermore, ISD measures will be used to show that the LP spectrum from this method is a better estimate than the summing beamformer used in Section 5.4.

5.7.4 Source Separation Techniques

In the previous section of this chapter, the AVS channels were processed using the well known source separation technique of fast ICA and convolutive fast ICA. The results showed that for noise corrupted speech signals, an improvement in terms of the perceptual quality was achieved. Although there was an improvement in the perceptual quality a significant amount of noise still remained in the processed signals. Here, the outputs of the fast ICA algorithm will be used to form the LP spectrum that is needed by the filter described in Section 5.4 and the outputs from the fast ICA will be filtered using those LP spectra. Here, only one ICA output is chosen as the other outputs correspond to noise. The correct ICA output is chosen based on the PESQ MOS score. In an automated system this is a drawback of using the fast ICA algorithm as it is difficult to identify the actual source from the outputs. In addition to the Fast ICA technique, a DOA based source separation technique was discussed in Section 5.6. Here this technique will also be used to enhance the noise corrupted speech.

5.7.5 Multivariate Auto-regression (MVAR)

The MVAR method was briefly discussed in Chapter 2; here the MVAR method for obtaining multichannel LP coefficients from recordings of an AVS will be presented. Let the AVS outputs be expressed as:

$$\boldsymbol{o(n)} = [x(n) \quad y(n)] \tag{157}$$

The linear prediction for vector ***o(n)*** which is 1×2 and can be expressed similar to (152) as:

$$\hat{\boldsymbol{o}}(n) = \sum_{i=1}^{p} \boldsymbol{a}_i \boldsymbol{o}(n-i) + \boldsymbol{e}\,(n) \tag{158}$$

where p is the prediction order and $\boldsymbol{a}_i$ is the i^{th} prediction matrix. The derivation of the multichannel prediction coefficients is based on the Levinson-Wiggins and Robinson (LWR) algorithm.

The difference between (156) and (158) is $\boldsymbol{a_i}$ is a prediction matrix not a prediction coefficient. The output of the MVAR process gives p prediction matrices, where each matrix is of size 2×2, and hence the total size of output is $2 \times 2P$. Matrix $\boldsymbol{a_i}$ can be expressed as [107]:

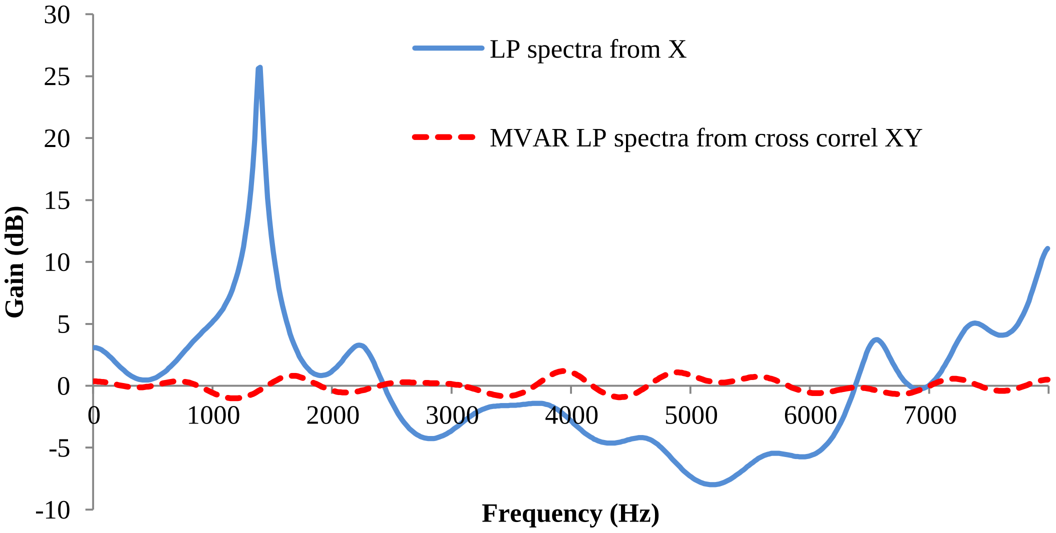
$$\boldsymbol{a_i} = \begin{bmatrix} a_i & c_i \\ b_i & d_i \end{bmatrix} \tag{159}$$

where a_i, d_i are equivalent to prediction coefficient from the autocorrelation R_{xx}, and R_{yy} of the channels x and y. The remaining prediction coefficients c_i, b_i are from the cross correlation of the channels [107]. Hence, the output of the MVAR algorithm contains the prediction coefficients due to auto and cross correlations of the channels.

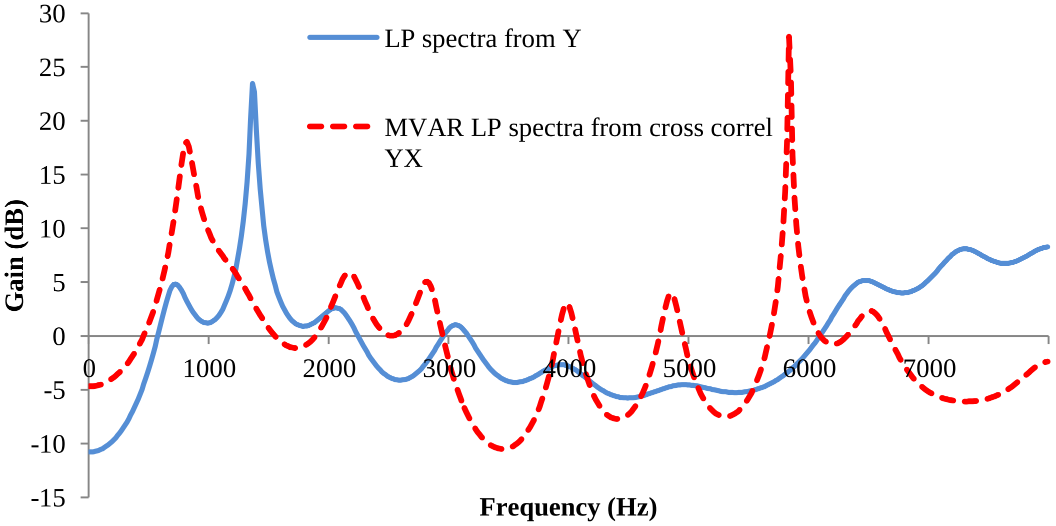
To investigate the outputs from the MVAR algorithm, an AVS recording of vowel *"a"* from a female speaker is analyzed. The LP spectra from the autocorrelation and cross correlation coefficients are shown in Figure 98 and Figure 99. The spectra formed from the LP coefficients of the autocorrelation in MVAR are very similar as that of the LP spectra from a single channel, the LP spectra from the cross correlation coefficient from the MVAR does not resemble the LP spectra of the single channel x or the y components. Hence, if the MVAR algorithm were to be used, the filter based on the LP spectra will not be accurate from the cross correlation. Furthermore, if the LP coefficients from the MVAR algorithm were to be used then individual channels have to be filtered separately. In addition to these issues, the stability of the MVAR algorithm suffers under some conditions. These include: when the block toplitz matrix used in the LWR algorithm tends be singular; when a signal is 0 or close to 0, or when signals are exactly the same [107]. This is a serious issue for the AVS as when a source it at 0 degrees in azimuth the output of the x channel is approximately 0 and when the source is at 45 degrees in azimuth both x and y channels are exactly the same.

5.7.6 Measuring the accuracy of the LP spectrum of a signal based on AVS channels

The accuracy of the LP spectra can be measured using two different approaches as discussed in 5.4.3. In Section 5.4.3 the LSD measure was used to show that the output of the summing beamformer for the AVS gradient channels were a closer match to the LP spectra of the clean speech compared to the output of the omni direction sensor or when the omni-directional sensor output was used in the summing beamformer. Here, the ISD which was described in detail in Section 2.9 will be used to investigate the accuracy of the LP spectra from methods outlined in Section 5.8.1. The lower the ISD value the closer the processed signal is to the desired signal. Figure 100 shows the ISD measured for the all the algorithm that has been presented above with the exception of the MVAR algorithm.



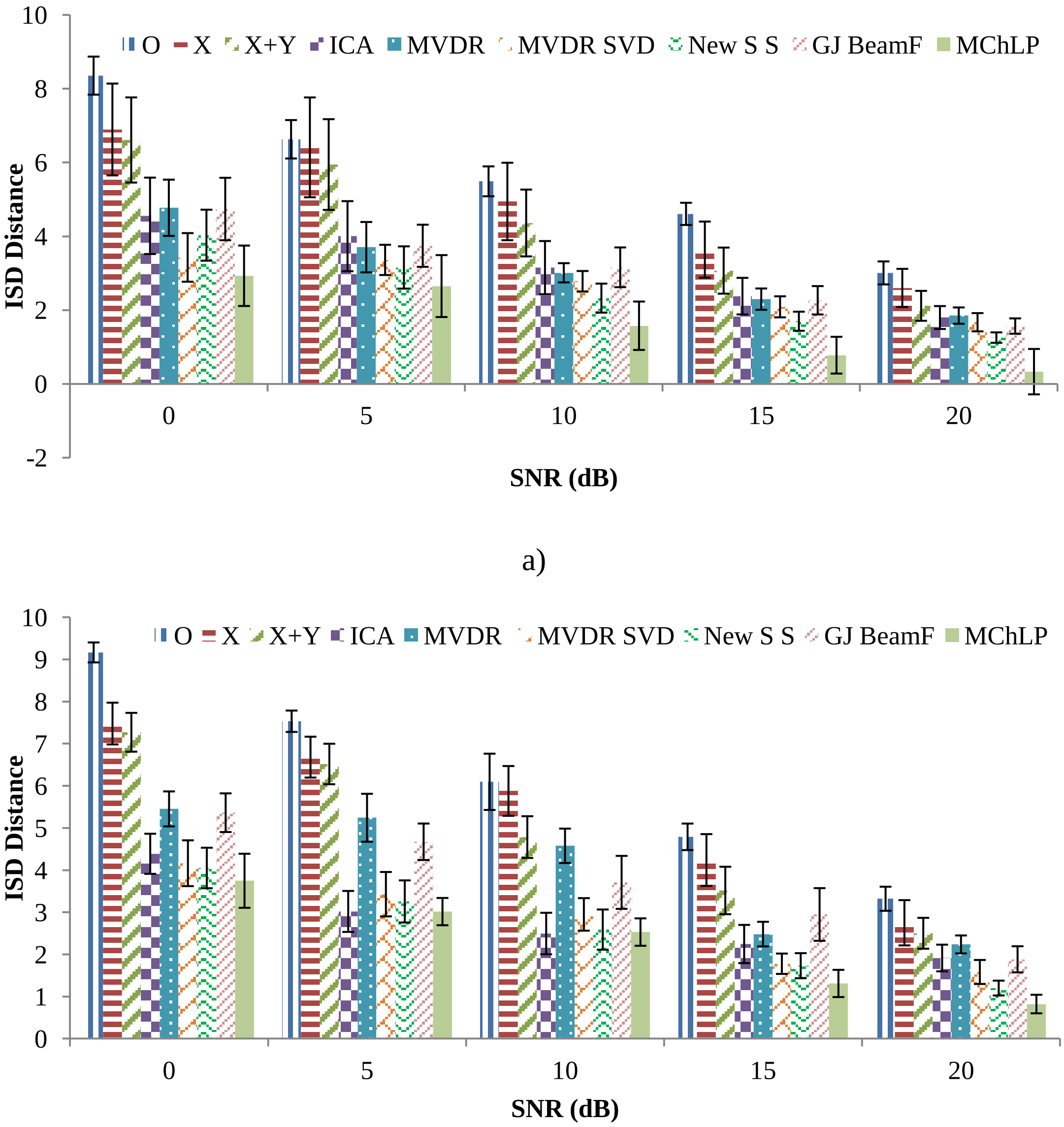
a) *x* channel



b) *y* channel

Figure 99: The LP spectra from autocorrelation Vs LP Spectra from Cross Correlation from MVAR

The results presented in Figure 100 are for the noise database described in Section 5.2.1 for different SNR levels. The signals are 10 sec long and each signal is processed using 20 ms frames with an overlap of 50%. The ISD is found for each frame and the results for ISD are averaged for all frames. The results presented in Figure 100 show that when an enhancement approach is applied to the AVS outputs the ISD is improved. Compared to the unprocessed channels, there is a considerable improvement using all algorithms. The enhanced MVDR, ICA, DOA based source separation algorithm all show lower values of ISD compared to the MVDR and the GJ



a)

b)

Figure 100: ISD measure for different algorithms (a) in anechoic (b) reverberant for different SNR’s (Error bars indicate 95% confidence intervals)

beamformer. The lowest values for ISD were obtained by the averaged autocorrelation function.

5.7.7 Experiments and Results

Experiments were performed to compare the performance using objective and subjective measures of all the algorithms after filtering with the perceptual filter. The perceptual filter has two parts as described in Section 5.4, which is pre-processing to

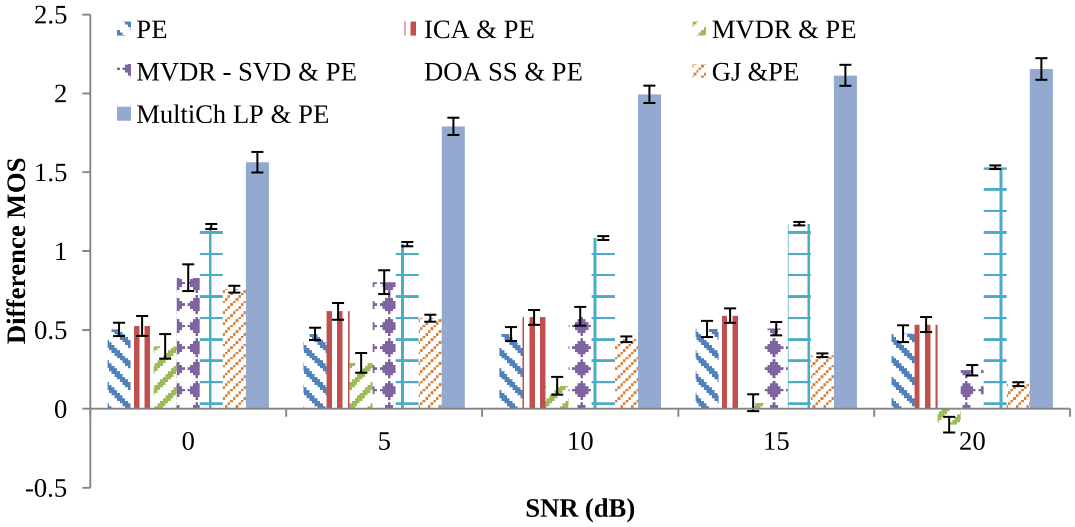


Figure 101: Difference MOS after perceptual filtering for anechoic recordings (Error bars indicate 95% confidence intervals)

obtain an accurate LP spectra and filtering the processed signal with the LP-spectra based perceptual post filter. In this experiment, the LP spectrum is obtained from the output of the different algorithms tested and those outputs are filtered using the perceptual filter.

5.7.8 Experimental setup

The database of speech and noise described in Section 5.2.1 are used here, with similar sampling rates, and conditions as those described in Section 5.4. The output from the filters were analysed using the techniques described in Section 5.4.

5.7.9 Experiments with Real Recordings

The results of the experiments shown in Figure 101 and Figure 102 are for average difference MOS for AVS outputs of different types of diffuse noise and for averaged SNR, for a target at 45 degrees in azimuth, filtered with different algorithms. In Figure 101 and Figure 102 the legend is interpreted as follows: MVDR – SVD is the enhanced MVDR Beamformer, DOA SS is the proposed source separation technique based on DOA estimation and GJ is the output form Griffiths and Jim beamformer and averaged autocorrelation function is indicated as MultichLP & PE. The PE at the end indicates Perceptual filtering. The following important conclusions can be drawn from the results.

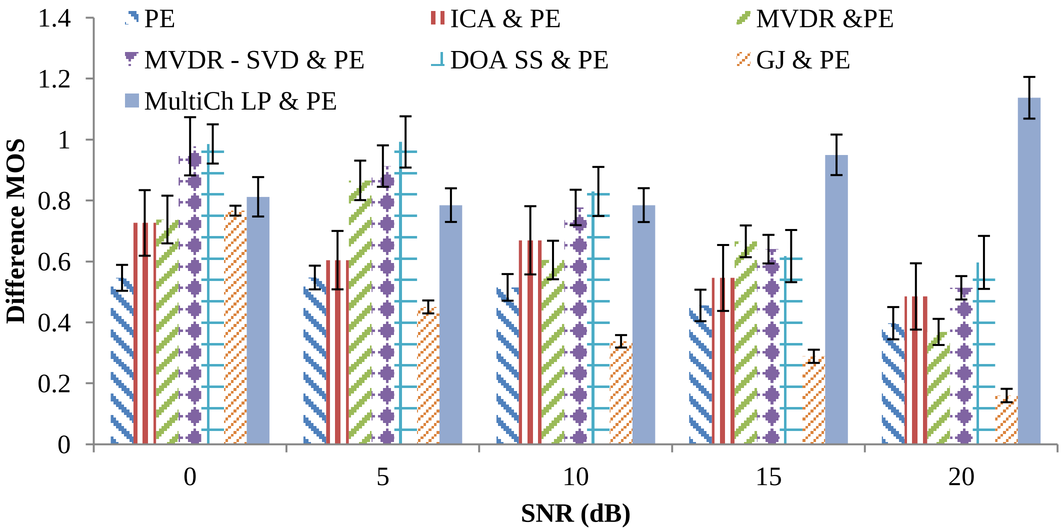


Figure 102: Difference MOS after perceptual filtering for reverberant recordings (Error bars indicate 95% confidence intervals)

a) The LP spectra obtained from source separation based on DOA estimation is a closer match to the clean source than the LP spectra obtained from fast ICA at higher noise levels.

b) Using the beamformers and the BSS algorithms to obtain the LP spectrum for the perceptual filtering improves the performance. But when compared with the multichannel LP, Multichannel LP performs better in anechoic conditions and at lower noise levels than combining the two filters. The difference in performance is due to over filtering when the beamformers and BSS algorithms are used. The over filtering introduces distortions which result in poor performance.

c) When the results of the ISD and PESQ are compared it is seen that those algorithm that showed lower values of ISD performed better in terms of PESQ. Hence it can be said that the ISD measure can be used as estimator for PESQ.

5.7.10 Results for Listening Test

The filters described in this chapter are used in obtaining the LP spectra and are used for filtering. The results presented here and shown in Figure 103 are for listening test conducted similar to Sections 5.2.2 and 5.3.7. For each algorithm, six files are used in the listening test.

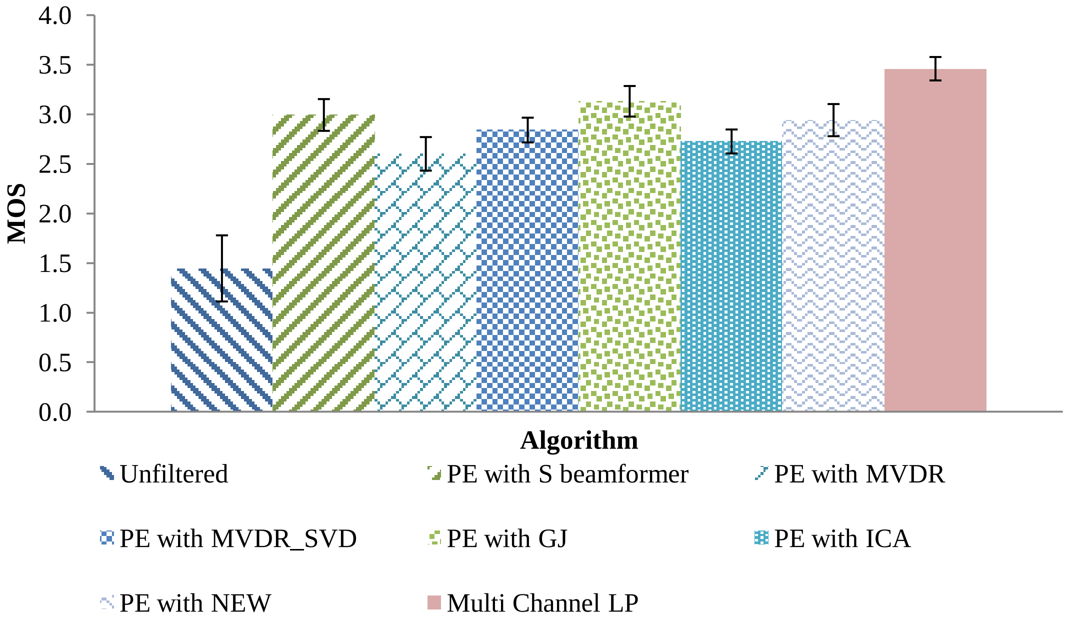


Figure 103: Results of listening test for different algorithms (Error bars indicate 95% confidence intervals)

The results show, compared to the summing beamformer used in Section 5.4 that all the other enhancement techniques except the multichannel LP from averaged autocorrelation did not produce any significant improvement. The averaged autocorrelation method for obtaining the LP coefficients produced an improvement of 0.5 MOS compared to all the algorithms used here. Furthermore, for most algorithms used in this experiment, it was found that by filtering the outputs of these algorithms with the perceptual filter resulted in a MOS score which is less that what was achieved without the perceptual filtering.

When the outputs are analyzed it was found the with all the beamformers and source separation algorithms the output of the perceptual filter were over filtered. Hence, even though the amount of noise removed is high, distortions were introduced. From these results and results of the PESQ and ISD results it can be concluded that for the multichannel perceptual filter the best mechanism to obtain the LP spectra is the averaged autocorrelation function.

5.7.11 Summary

The results presented in this section has shown that the by using multichannel speech enhancement techniques such as beamforming and source separation algorithms, it is possible to get a closer LP spectra to that of the clean signal from noise corrupted signals. But it was also identified that by using these algorithms with a perceptual LP based filter the output of the filter did not show any significant improvement and furthermore it was found that by using these algorithms distortions due to over filtering reduced the quality of the output. From the results presented in this section, it can be concluded that the best combination for obtaining the LP spectra and perceptual filtering is to use the averaged autocorrelation function to obtain a LP spectra and to use the summing beamformer for combining the channels for perceptual filtering.

5.8 Dereverberation of Speech Source using an AVS

The work presented so far in this chapter investigated speech enhancement using beamforming, perceptual filtering and BSS algorithms. In this section, speech enhancement using dereverberation will be presented. The effect of reverberation in speech and audio plays an important role in making the sound more natural, but when the amount of reverberation is too high it reduces the speech signal quality. Hence, it is important to find ways of removing reverberation. Reverberation mostly affects hands free applications such as teleconferencing, voice activated control of electronic equipment and automatic transcription systems. As discussed in Section 2.9.3, there are several methods that are proposed for dereverberation, but most of these methods rely on the accurate estimate of room impulse response or Acoustic Transfer Function (ATF), which is difficult to obtain accurately. In recent literature, dereverberation based on multichannel recording such as those from a microphone array has been reported. These multichannel techniques have two stages, which are:

a) A beamformer (e.g. DS Beamformer)

b) A single channel postfilter

Such two stage approaches, such as SMERSH cannot be used in real time applications due to the delays in the processing as they require multiple frames to be analysed.

In this work a multichannel dereverberation method is proposed which does not rely on the ATF, and performs the dereverberation in one stage using the DOA estimates from an AVS. Unlike other microphone arrays, the AVS with its directional pressure gradient sensors capture less reverberation compared to an omni-directional microphones, which provides an advantage where some amount of dereverberation is achieved at the microphone array output.

In this section the method, presented in Section 4.5 and 5.6 for DOA estimation and source separation will be used for enhancing reverberant speech signals. In chapter two the components of a reverberant soundfield were discussed. In a reverberant soundfield there are three components which are the direct component, early reflections and late reflections. The method described in Section 5.6 sorts the frequency components of the recordings from the gradient sensors which are clustered into different frequency components from different DOAs. In reverberant conditions, the direct component can be thought of as a source, and the two reflected components can be thought of as different sources from different DOAs. By sorting the frequency components into these three components the direct component can be obtained. The frame length used in Section 5.6 is 20 ms and as a result the algorithm in 5.6 can be used in real time for source separation. Here, the same algorithm will be used with a frame length of 10ms and hence real time dereverberation can be achieved.

The performance of the technique used in this section will be compared against the well known SMERSH algorithm. The SMERSH algorithm used here will be based on the multichannel implementation described in [108] and using the multichannel DYPSA algorithm implemented and provided by Mark R. P. Thomas of Imperial College, London, UK. The results presented in this section will be for two microphone arrays, and the AVS will be compared against the Core Sound TetraMic [175, 176] which is similar to a soundfield microphone but without the protective netting. The experimental setup is described next.

5.8.1 Experimental setup

The recording used in this section is made in room 4 m wide by 12 m long by 3 m high with concrete walls and only two doors. The ceiling and the floor are also concrete with little furniture such as chairs and table. The RT_{60} for the room was found to be 1.5 seconds. The database of speech sources described in Section 5.2.1 is used. Recordings are made with loudspeakers located 1,2,3,4 and 5 m from the array. The AVS is fixed such that the loud speakers are at zero degrees in azimuth to the array as shown in Figure 104. Recordings are made at 48 kHz and down sampled to 16 kHz.

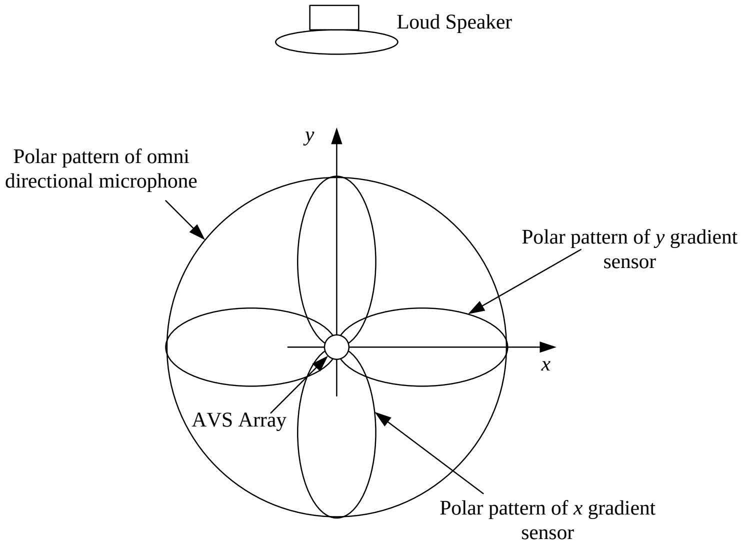

Loud Speaker
y
Polar pattern of omni directional microphone
Polar pattern of y gradient sensor
x
AVS Array
Polar pattern of x gradient sensor

Figure 104: Experimental setup for Reverberant recordings

5.8.2 The SRR for the Unprocessed Recordings from the AVS and Core Sound TetraMic

In this section the amount of reverberation captured by the individual channels of the AVS and the TetraMic will be investigated. These microphone arrays have similar polar responses as shown in Figure 104. The AVS and the TetraMic are placed in front of the loud speaker such that the array is at zero degrees in azimuth relative to the loud speaker as described above and shown in Figure 104. The SRR ratio as described in [29] is used in this work. The SRR between the original clean recording and o, x and y channels of the AVS and TetraMic are shown in Figure 105 a) and b).

From the results is seen that the amount of reverberation captured by the x gradient sensor is less than that of the y gradient sensor. When compared with the y gradient sensor, the omni-directional sensor captures less reverberation. The reason for the high amount of reverberation captured buy the y sensor is, due to its polar pattern, since it is placed perpendicular to the direction of the direct component from the loud speaker, very little of the direct component is captured by the y sensor. In contrast, the x sensor is parallel to the loudspeaker and so has maximum capture of direct component.

The bulk of the signal that is captured by the y sensor is the early reflections and late reflections from the walls in the x direction. The x sensor on the other hand captures the direct component and the majority of the reflections captured by the x sensor are those from the y direction. The omni-directional sensor captures the direct component as well as all the reflections from all the direction. From the results in Figure 105 (b) it can be seen that compared to the AVS, the TetraMic captures more reverberation from all the channels. As described in Section 2.8.6 the Soundfield microphones are arranged in a tetrahedron configuration and the directional components are formed using the outputs of the four microphones capsules. Hence it captures more reflections than an AVS which contains directional microphones.

In addition to the difference in the amount of reflection captured by the different microphone arrays and different channels, the amount of reflections captured by both the arrays increase as the separation between the array and loud speaker increase. This is expected since, the distance between the source and microphone increase, there is time for more reflections to occur and the amount of reverberation increases.

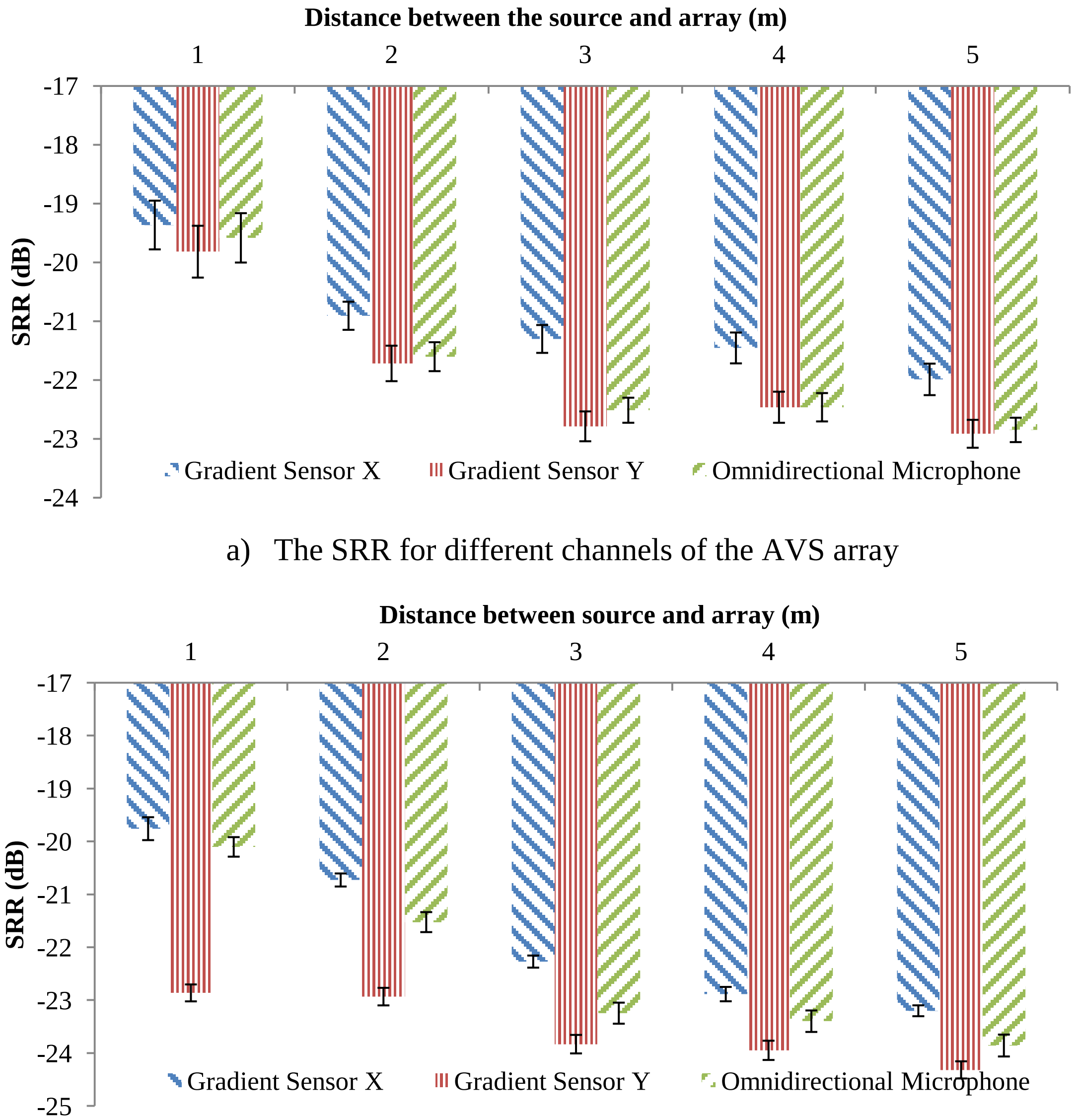
5.8.3 Dereverberation

The method used in this section for dereverberation is exactly the same technique described in Section 5.6, with different thresholds and scaling factors. The threshold used in the Sections 4.5 and 5.6 were for two and three sources is 0.6. Here, since only the direct component is desired, the threshold can be increased. From Figure 60 it can be seen that by increasing the threshold from 0.6 to 0.75 the number of sources detected is approximately 1, hence in this part the threshold is set to 0.75. In addition to the threshold for determining the number of sources, the scaling factor described in Section 5.6.3 is also lowered to 0.1 such that the amount of reverberation added back is minimised. The processing of the recordings is performed with a frame length of 10ms, and an overlap of 50% with a 2048 point FFT.

The result of the dereverberation using the DOA method is presented in Figure 106 a) and b) for the AVS and the Tetra Mic respectively. From the results it can be seen the proposed method performs much better than the Multichannel SMERSH (MC SMERSH) algorithm for both arrays.

The results show that on average the proposed method has an improvement in terms of difference in gain between the unprocessed x sensor recordings to the output of the proposed algorithm of 1.5 dB for a separation of 1m and increases in difference in gain to 2.6 dB for 5m.

For the MC SMERSH algorithm the improvement in gain at 1m is 0.5 dB and the difference in gain increases to 0.8dB at 5m for the AVS. In the case of the Tetra



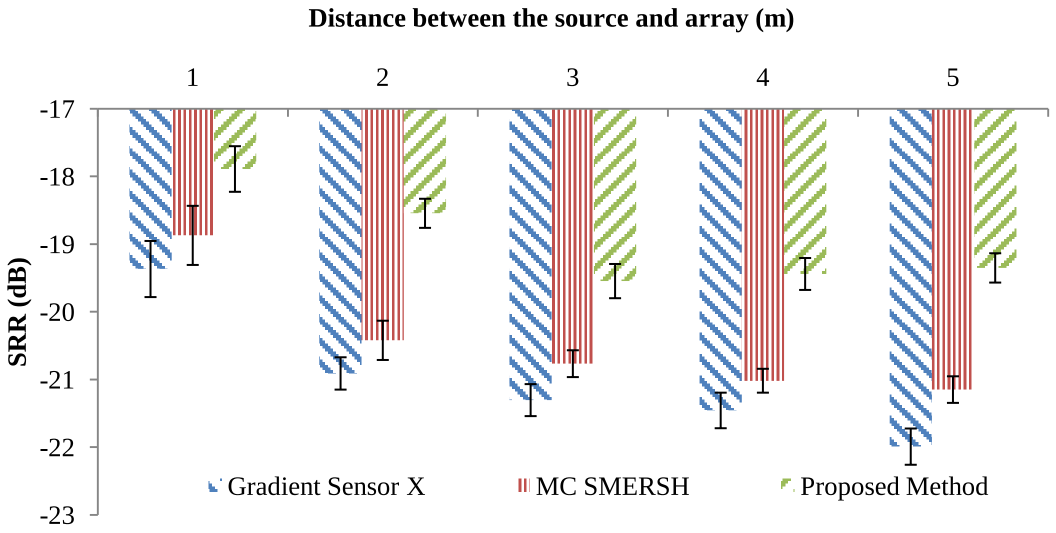
a) The SRR for different channels of the AVS array

b) The SRR for different channels of the Core Sound TetraMic

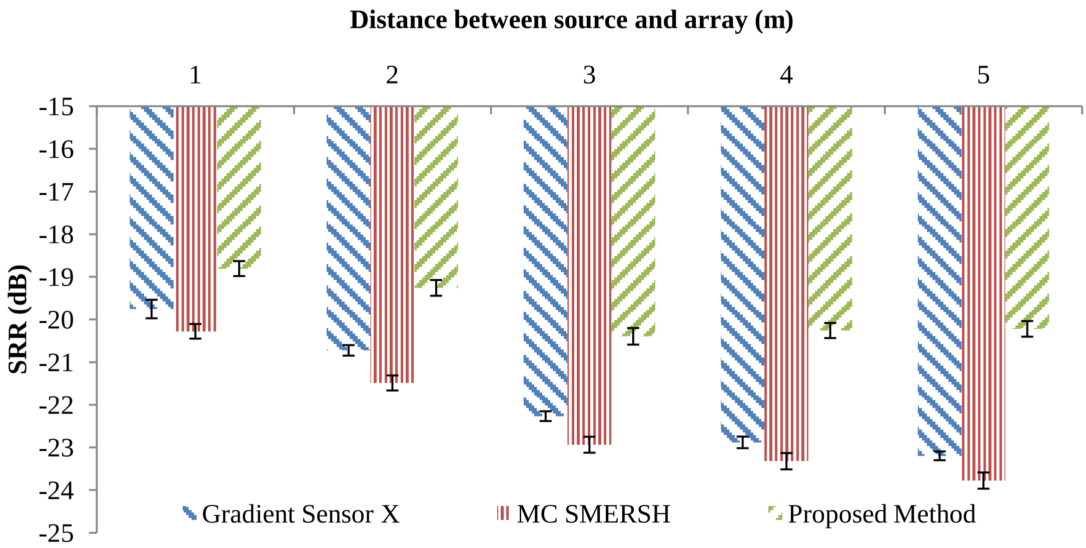
Figure 105: The SRR for Different Channels of the AVS and TetraMic arrays (Error bars indicate 95% confidence intervals)

Mic. the improvement in terms of the difference in gain for 1m is 0.9 dB and the difference increases to 2.9 dB for the proposed method.

For the MC SMERSH algorithm applied to the Tetra Mic. outputs, the improvement obtained for 1m is 0.5 and for 5m it is seen that the performance stayed the same at 0.58 dB. As mentioned in the previous section the amount of reverberations picked up by the microphone arrays increase as the distance between the source and array increase.



a) The results for SRR for MC SMERSH and Proposed Method for AVS



b) The results for SRR for MC SMERSH and Proposed Method for TetraMic

Figure 106: The results of Dereverberation using MC SMERSH algorithm and proposed method for AVS and TetraMic (Error bars indicate 95% confidence intervals)

The MC SMERSH algorithm was designed for use with a distributed arrays and the DS beamformer algorithm was used in obtaining the LP residual, which was needed to identify the glottal closure instances. But since one channel of the AVS and TetraMic contained only reverberation, by including that channel in the processing increased the errors in the output of the MC SMERSH algorithm.

To correct this error the channel with only reverberations can be excluded, but this is not a practical solution, since the direction of the source may change and the channel that needs to be excluded may change and not all the DOAs will have channels with dominant reverberations. In the case of the proposed method, only the components that contain the DOA of the direct component are used for both channels, hence is more suited to an array such as the AVS. From the results it can be clearly seen that although the MC SMERSH algorithm performs well with other microphone arrays, for an AVS or a similar co-located microphone array such an approach is not suitable.

The results presented so far are for a frame length of 10 ms; here the frame length is increase from 10ms to 500ms and to 1s. In most dereverberation algorithms to remove late reflections the frame length has to be increased to match the reverberation time. Here, for the proposed technique, the effect of increasing the frame length will be investigated. The results of increasing the frame length for processing the AVS recording are presented in Figure 107. From the results it can be seen clearly that that by increasing the frame length from 10ms to 1s the performance of the proposed method decreases in terms of SRR. This is expected, since the frame length increases, the number of reflections in the frame from the y direction increases; and the algorithm assumes that these components are part of the direct component. The effect of increasing the frame length is to include delayed and added components of the same speech section due to the reflected components.

5.8.4 Summary

The result presented in this section has shown that proposed speech source separation algorithm in Section 5.6 can be used for dereverberation. From the results, it seen that by using proposed source separation algorithm based on clustering DOAs of different frequency components, it is possible to dereverberate speech in adverse conditions ($RT_{60} > 1s$) without using the ATF, and with frame lengths of 10ms with recording from an AVS and a Core Sound Tetra Mic. These results have also shown that multichannel recordings from a co-located array can be used for speech dereverberation.

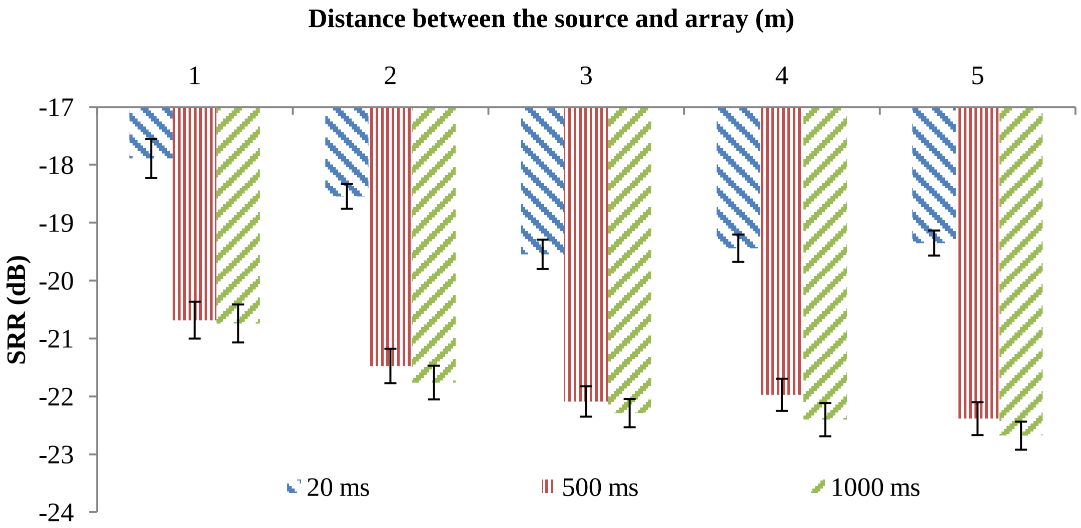


Figure 107: Results of SRR for increasing the frame length of the propose method for the recordings of the AVS (Error bars indicate 95% confidence intervals)

The results presented here has also shown that due to the directional sensors of the AVS, the array is capable of capturing signals with reduced reverberation compared with an omni-directional sensor. The results presented in this chapter have shown different ways of enhancing noise corrupted speech signals.

5.9 Conclusions

In this chapter, five different methods for enhancing the outputs of the AVS corrupted by noise have been presented. The results presented showed that by taking advantage of the directional characteristics of the AVS array, the enhancement algorithms produced improvements in the quality of noise corrupted speech signals. The result presented in this chapter has also shown that the AVS array can be used for dereverberation and source separation. It was shown that due to the direction characteristics of the array the recordings from the AVS have reduced reverberation.

The results presented in this chapter are summarised in Table 5, from which it can be concluded that for the AVS array the enhancement algorithms that produces the best performance are the perceptual filter based on the average autocorrelation function presented in Section 5.7, source separation algorithm presented in Section 5.6, and the fastICA work presented in Section 5.5. These enhancement algorithms showed the highest values for difference MOS for PESQ and listening tests. The results also

	Difference MOS	
	PESQ	Listening Test
Summing Beamformer	0.19	0.11
MVDR Beamformer	0.23	0.30
MVDR - SVD Beamformer	0.59	1.74
Griffiths and Jim Beamformer	0.40	1.77
Perceptual Filter DS Beamformer	0.49	1.63
Perceptual Filter Average Autocorrelation	1.41	1.98
ICA	0.74	1.75
Proposed DOA Based Source Separation	0.78	2.09

Table 5: Comparison of all the enhancement algorithms presented in Chapter 5

showed that the proposed MVDR-SVD beamformer and the GJ beamformer showed significant improvements in MOS for listening test. The results show that for the AVS array speech enhancement algorithms based on perceptual filtering and source separation work better than beamforming algorithms. The next chapter will present the conclusions and summary of the work that has been presented in this thesis.

Chapter 6 Conclusions and Future Research

6.1 Introduction

The work presented in this thesis is based on an in air-AVS for speech signal processing. The work presented in this thesis is unique as speech signal capturing and processing with a co-located microphone array such as the AVS has not been done before. Here, the array design was modified to suite in-air applications such as speech enhancement and DOA estimation. In addition to the design, different types of signal processing for an AVS array was looked at which included, beamforming, filtering, dereverberation and source separation. It was shown that due to the directional sensors on the array, advantages in terms of better performance were achieved for all the algorithms. The next section in this chapter will look at the results obtained in this work and the future research areas for the AVS.

6.2 Design of the AVS

The work presented in Chapter 3 has shown that the design of the AVS array plays an important role in the quality of recorded signals and obtaining accurate measurements of the directional information from the directional sensor on the AVS array. It was shown that there is direct link between the accuracy of the DOA estimation and the placement of the sensors on the array and by adjusting the positions of the sensors on the array more accurate results for DOA estimations were obtained.

It was shown that the accuracy with which the directional information from the pressure gradient microphones depended on an artificial increase of the surface area of the microphone and front to back separation of the microphone due to the placement of the microphones and structure holding the microphones in place. By modifying the design such that the artificial increase due to these factors described above are reduced, the DOA estimates of source with an accuracy of 2 degrees for a source at 1m can be achieved.

6.3 DOA Estimation

Chapter 4 looked at different methods of obtaining a DOA estimate from the AVS array in realistic conditions. The well known MUSIC algorithm for DOA estimation for microphone arrays and an intensity based time domain algorithm and a frequency domain algorithm for an AVS array were presented. It was shown that all the algorithms used showed accurate results for monotone signals, but only the MUSIC algorithm and the frequency domain intensity algorithm could be used for speech signals. The results presented in Chapter 4 showed that using these two algorithms, DOA estimates for stationary and mobile sources could be obtained for a single frame of 10 ms with an accuracy of approximately 1.5 degrees for stationary sources and approximately 4 degrees for mobile source. The results from the AVS were compared with the results from a Soundfield microphone which has similar directional characteristics as the AVS. When the DOA estimates from the Soundfield microphone is analysed it is seen that accuracy of the DOA estimates are lower than that of the AVS with an accuracy of approximately 5 degrees for stationary source and an accuracy of 30 degrees for moving sources.

From the results it was found that the MUSIC algorithm performed better for both stationary and mobile sources in comparison to the frequency domain intensity algorithm and the performance of the AVS is much better than a Soundfield microphone for DOA estimation.

The DOA estimation for multiple consecutive speech sources was investigated in Section 4.5 of Chapter 4. It was found that due to the complexity in the outputs of the DOA estimation from both MUSIC and frequency domain intensity algorithms, data clustering techniques had to be used to obtain DOA estimates. A method based on hierarchical and partitional clustering combined was presented to sort the DOA from both algorithms, where the sorting for the MUSIC algorithm was performed for multiple frames and sorting was performed on a frame by frame basis and for multiple frames for intensity based frequency domain algorithm. The results showed that DOA estimates with an accuracy of approximately 5 degrees could be obtained from the frequency domain intensity based algorithm, but with the MUSIC algorithm it was not possible to obtain a statistically valid result. The DOA estimation for the recordings from the Soundfield microphone for multiple sources showed an accuracy of 13 degrees could be achieved. The results showed that while the MUSIC algorithm was better for estimating DOAs for a single source, the frequency domain intensity based algorithm is much more suited to multiple sources.

6.4 Speech Enhancement

The work on speech enhancement was presented in Chapter 5, where five different methods for speech enhancement were discussed. Here the results for different algorithms will be summarised.

6.4.1 Results of Speech Enhancement Using Beamformers

The work in beamforming showed the different beamformers that could be used with an AVS and it was shown that using the directional characteristics of the AVS array an accurate estimate of the noise covariance matrix could be obtained and when used with the MVDR beamformer, improved performance over the original MVDR beamformer could be achieved.

The beamformers when used for enhancement of noise corrupted speech recorded by an AVS showed there is an improvement of 0.1 and 0.2 MOS for the summing beamformer over the original in anechoic and reverberant conditions respectively. The conventional MVDR beamformer recorded an improvement of 0.2 and 0.3 MOS over unprocessed recording for anechoic and reverberant conditions and the enhanced MVDR beamformer recorded an improvement of 0.6 MOS for both anechoic and reverberant conditions over the unprocessed recording. The GJ beamformer also showed improvement in MOS of 0.4 for anechoic and reverberant conditions.

The results of listening tests for the beamformers showed that listeners ranked the unprocessed, summing and MVDR beamformers as bad and the enhanced MVDR beamformer and GJ beamformer as fair, which a considerable improvement in terms of MOS.

6.4.2 Results for Perceptual Multichannel Filter

The results from previous section has shown that by applying a beamformer to the AVS outputs, noise corrupted speech can be enhanced. In Section 5.4 a multichannel perceptual filter is proposed for enhancing of noise corrupted speech recorded with an AVS. The original implementation of the perceptual filter was a single channel algorithm. In this work the original implementation is modified to take advantage of the directional sensors of the AVS. The proposed method requires obtaining the LP spectra of the clean speech for the filtering process. It is shown that by using the directional sensors of the AVS, a much better estimate of the LP spectra of clean speech can be obtained from the noise corrupted speech. It is also shown that when the omni-directional sensor is included in the filtering process, the results are worse than using only the directional sensors. The results show an improvement of 0.3 and 0.2 MOS for anechoic and reverberant case over the MVDR beamformer and 0.1 and 0.14 MOS improvement over the GJ beamformer in anechoic and reverberant case. From the listening test it is seen that the perceptual filter shows an improvement of 1.6 MOS over unprocessed recording, where as the MVDR beamformer only shows a 0.1 MOS improvement over the unprocessed recordings. The proposed method has shown that by using the directional components of the AVS in the enhancement, better results for enhancement are obtained.

6.4.3 Results for Speech Enhancement Using FastICA

In Section 5.5 speech enhancement using source separation based on Fast ICA algorithm is discussed. Fast ICA algorithm requires the number of recordings to be equal or less than the number of sources and the recordings have to be statistically independent of each other. In Section 5.5, it is shown that due to the directional components of the AVS recording there is a difference in kurtosis and mutual information measure between the channels of the AVS which allows ICA to be applied to the recordings made with AVS. Comparisons made between different microphone arrays show that when the microphone arrays have directional sensors the performance of ICA is better than the arrays that have omni-directional sensors. The results from the work presented in this section have shown the significance of directional sensors in the AVS array.

6.4.4 Results for Speech Enhancement by Source Separation

The frequency domain intensity based method used in Chapter 4 for DOA estimation of multiple sources using data clustering is extended to source separation. The frequency domain intensity based DOA estimation methods enables a DOA estimate for each time frequency component estimated by an FFT. Since the formant frequencies of different speakers differ, using the DOAs obtained for different FFT points, the frequency components belonging to different sources can be separated. The results show that when compared with unprocessed recording, the proposed method produces an improvement of 1 MOS LQO for two speakers and an improvement of 1.1 MOS LQO for three speakers. Overall the improvements were better than the ICA algorithm.

6.4.5 Results for Obtaining Accurate LP Spectra for Perceptual Filtering

In Section 5.7 different methods that can be used for obtaining an accurate LP spectra was proposed. The speech enhancement algorithms described in Chapter 5 were used to obtain an accurate estimate of the LP spectra. In addition to the speech enhancement algorithms the methods that could be used to obtain a multichannel LP spectra was also shown. The results showed that by using the different speech enhancement algorithms gave LP spectra which was a better estimate than that obtained from the summing beamformer, but the output of the perceptual filter did not show a significant improvement due to distortion from over filtering. The best results was obtained by using the multichannel LP spectra obtained from the average autocorrelation function, which showed an improvement of 0.5 MOS over the results obtained in Section 5.4.

6.4.6 Results for Speech Enhancement by Dereverbration

The source separation algorithm described in Section 5.6 is used in dereverbration of speech corrupted by reverberation. It was shown that by using the source separation technique to separate the direct component, the reverberant speech signals were enhanced. Furthermore it was shown that the directional components of the AVS array contained less reverberation compared to the omni-microphone. The results showed that the proposed method showed an improvement in terms if SRR of 1.5dB over the unprocessed recording at 1m and 2.6 dB at 5m. Comparisons with the MC SMERSH algorithm showed that overall the proposed method showed better improvements.

6.5 Future Research Areas

The work presented in this thesis has shown that by using an AVS to capture speech signals, accurate DOA estimation and speech enhancement can be achieved. In this work, it was shown that by utilizing the directional information from the AVS channels source separation and dereverbration can be achieved. Here, the array used for these experiments is a two dimensional array. Future research can be extended to three dimensional arrays by including the *z* sensor.

One of the most accurate particle velocity sensors is the hotwire anemometer, which has so far not been used for capturing speech signals due to the low SNRs. In this work, it was shown that the DOA estimates from the directional sensors from the AVS gives a DOA for each frequency components. This can be extended such that this relation between the DOA and frequency component from a hotwire anemometer can be used to separate the sources using a recording from an omni-directional microphone which has a better SNR.

The AVS array used in this work is extremely compact compared with other microphone arrays, but it is still too large for mobile devices such as mobile phones. In recent years several miniature omni-directional microphones have been introduced. These microphones are about 3 by 3 mm and could be used to form an AVS which is smaller and much more suited to small electronic devices such as mobile phones.

6.6 Conclusions and Summary

The work presented in this thesis has shown that an AVS, which is a compact co-located microphone array, can be used for capturing and processing speech signals. It was shown that the directional sensors of the AVS array provided advantages in DOA estimation, dereverbration and speech enhancement. The results showed that compared to a Soundfield microphone, the performance of the AVS array is better in terms of

DOA estimations and improvements in enhancements of noise corrupted speech in terms of subjective and objective tests.

A beamforming technique which utilizes the directional components of the AVS for obtaining an accurate noise covariance matrix was proposed. In addition to the beamformer, a perceptual filter which utilizes the directional information to obtain an accurate LP spectrum was also proposed. The work in this thesis showed that due to the directional components of the AVS array, source separation based on the DOAs from the array and the FastICA algorithm could be used for enhancement of noise corrupted speech. The work in this thesis has shown that using the AVS a significant improvement to many speech signal processing applications can be achieved.

References

[1] M. Shujau, C. H. Ritz, and I. S. Burnett, "Designing Acoustic Vector Sensors for localisation of sound sources in air," presented at the *17th European Signal Processing Conference (EUSIPCO 2009)*, Glasgow, Scotland., 2009.

[2] M. Shujau, C. H. Ritz, and I. S. Burnett, "Using in-air Acoustic Vector Sensors for tracking moving speakers," in *Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference on*, 2010, pp. 1-5.

[3] M. Shujau, C. H. Ritz, and I. S. Burnett, "Separation of Speech Sources Using An Acoustic Vector Sensor," presented at the *2011 IEEE internatinal Workshop on Multimedia Signal Processing (MMSP)*, Hanzhou, 2011.

[4] M. Shujau, C. H. Ritz, and I. S. Burnett, "Linear Predictive Perceptual Filtering For Acoustic Vector Sensors: Exploiting Directional Recordings For High Quality Speech Enhancement," presented at the *Acoustic Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, Praque, 2011.

[5] M. Shujau, C. H. Ritz, and I. S. Burnett, "Speech enhancement via separation of sources from co-located microphone recordings," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 137-140.

[6] M. Hawkes and A. Nehorai, "Effects of sensor placement on acoustic vector-sensor array performance," *Oceanic Engineering, IEEE Journal of,* vol. 24, pp. 33-40, 1999.

[7] A. Nehorai and E. Paldi, "Acoustic vector-sensor array processing," *Signal Processing, IEEE Transactions on,* vol. 42, pp. 2481-2491, 1994.

[8] A. Nehorai and E. Paldi, "Vector-sensor array processing for electromagnetic source localization," *Signal Processing, IEEE Transactions on,* vol. 42, pp. 376-398, 1994.

[9] M. E. Lockwood and D. L. Jones, "Beamformer performance with acoustic vector sensors in air," *The Journal of the Acoustical Society of America,* vol. 119, pp. 608-619, 2006.

[10] J. E. White, "Directional sound detection," ed: US Patent 2,982,942, 1961.

[11] D. Lindwall, "Marine Seismic Surveys with Vector Acoustic Sensors," *American Geophysical Union*, Fall Meeting 2006.

[12] M. Hawkes and A. Neharai, "Hull-mounted acoustic vector-sensor processing," presented at the Proceedings of the *29th Asilomar Conference on Signals, Systems and Computers* (2-Volume Set), 1995.

[13] M. Hawkes and A. Nehorai, "Acoustic vector-sensor processing in the presence of a reflecting boundary," *Signal Processing, IEEE Transactions on,* vol. 48, pp. 2981-2993, 2000.

[14] M. Hawkes and A. Nehorai, "Bearing estimation with acoustic vector-sensor arrays," *AIP Conference Proceedings,* vol. 368, pp. 345-358, 1996.

[15] M. Hawkes and A. Nehorai, "Wideband source localization using a distributed acoustic vector-sensor array," *Signal Processing, IEEE Transactions on,* vol. 51, pp. 1479-1491, 2003.

[16] R. Hickling, W. Wei, and R. Raspet, "Finding the direction of a sound source using a vector sound intensity probe," *The Journal of the Acoustical Society of America,* vol. 93, p. 2357, 1993.

[17] R. Raangs and W. Druyvesteyn, "Sound source localization using sound intensity measured by a three dimensional PU-probe," *PREPRINTS-AUDIO ENGINEERING SOCIETY,* 2002.

[18] P. K. T. Wu, C. Jin, and A. Kan, "A Multi-Microphone Speech Enhancement Algorithm Tested Using Acoustic Vector Sensors," *International Workshop on Acoustic Echo and Noise Control (2010),* 2010.

[19] G. S. Kino, "Acoustic waves," *Egewood Cliffs, NJ: Prentice-Hall, Inc,* 1987.

[20] D. M. Howard and J. Angus, *Acoustics and psychoacoustics*: Focal press, 2009.

[21] F. P. Mechel and P. J. Morris, "Formulas of acoustics," *The Journal of the Acoustical Society of America,* vol. 115, p. 941, 2004.

[22] C. H. Ritz, Schiemer, G., Burnett, I., Cheng, E., Lock, D., Narushima, T., Ingham, S., Wood Conroy, D.,, "An Anechoic Configurable Hemispheric Environment for Spatialised Sound," presented at the *Australasian Computer Music Conference,* 2008.

[23] H. E. de Bree, W. F. Druyvesteyn, E. Berenschot, and M. Elwenspoek, "Three-dimensional sound intensity measurements using Microflown particle velocity

sensors," *Micro Electro Mechanical Systems*, MEMS 1999, 12th *IEEE International Conferrence on*, 1999, pp. 124-129.

[24] H.-E. de Bree, "Add-On Microflown for a High-End Pressure-Gradient Microphone," *Audio Engineering Society Convention 109*, 2000.

[25] D. Havelock, S. Kuwano, and M. Vorländer, *Handbook of signal processing in acoustics* vol. 1: Springer Verlag, 2008.

[26] J. Eargle, *The microphone book*: Focal Pr, 2004.

[27] D. Davis and E. Patronis, *Sound system engineering*: Focal Press, 2006.

[28] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*: Springer, 2010.

[29] P. Naylor, *Speech dereverberation*: Springer, 2010.

[30] I. McCowan, "Microphone arrays: A tutorial," *Queensland University, Australia*, pp. 1-38, 2001.

[31] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*: Springer, 2010.

[32] D. G. Manolakis, V. K. Ingle, S. M. Kogon, and I. ebrary, *Statistical and adaptive signal processing: spectral estimation, signal modeling, adaptive filtering, and array processing*: Artech House, 2005.

[33] M. van der Wal and Start, "Design of Logarithmically Spaced Constant-Directivity Transducer Arrays," *J. Audio Eng. Soc*, vol. 44, pp. 497-507, 1996.

[34] Y. Tamai, S. Kagami, Y. Amemiya, Y. Sasaki, H. Mizoguchi, and T. Takano, "Circular microphone array for robot's audition," in *Sensors, 2004. Proceedings of IEEE*, 2004, pp. 565-570 vol.2.

[35] E. Hulsebos and Schuurmans, "Circular Microphone Array for Discrete Multichannel Audio Recording," *Audio Engineering Society Convention 114*, 2003.

[36] A. Karbasi and A. Sugiyama, "A new DOA estimation method using a circular microphone array," *Presented at European Signal Processing Conference*, EUSIPCO 2007, Poznan, Poland, 3-7 September 2007, pp. 778–782.

[37] H. Hacihabiboglu, E. De Sena, and Z. Cvetkovic, "Design of a Circular Microphone Array for Panoramic Audio Recording and Reproduction: Microphone Directivity.", *Audio Engineering Society Convention 128*, May 2010

[38] B. Rafaely, "Analysis and design of spherical microphone arrays," *Speech and Audio Processing, IEEE Transactions on,* vol. 13, pp. 135-143, 2005.

[39] Z. Li and R. Duraiswami, "A robust and self-reconfigurable design of spherical microphone array for multi-resolution beamforming," presented at the *Acoustic Speech and Signal Processing (ICASSP)*, 2005 IEEE International Conference on,2005, pp. iv/1137-iv/1140 Vol. 4.

[40] B. N. Gover, J. G. Ryan, and M. R. Stinson, "Microphone array measurement system for analysis of directional and spatial variations of sound fields," *The Journal of the Acoustical Society of America,* vol. 112, p. 1980, 2002.

[41] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 24, pp. 320-327, 1976.

[42] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," presented at the *Acoustic Speech and Signal Processing (ICASSP)*, 1997 *IEEE International Conference on*, 1997, pp. 375-378 vol. 1.

[43] A. Brutti, M. Omologo, and P. Svaizer, "Comparison Between Different Sound Source Localization Techniques Based on a Real Data Collection," in *Hands-Free Speech Communication and Microphone Arrays, 2008. HSCMA 2008*, 2008, pp. 69-72.

[44] Q. Bo, Z. Heng, F. Qiang, and Y. Yonghong, "Subsample time delay estimation via improved GCC PHAT algorithm," in *Signal Processing, 2008. ICSP 2008. 9th International Conference on*, 2008, pp. 2579-2582.

[45] L. Sangmoon, P. Youngjin, and P. Youn-sik, "Cleansed PHAT GCC based sound source localization," in *Control Automation and Systems (ICCAS), 2010 International Conference on*, 2010, pp. 2051-2054.

[46] K. Byoungho, P. Youngjin, and P. Youn-sik, "Analysis of the GCC-PHAT technique for multiple sources," in *Control Automation and Systems (ICCAS), 2010 International Conference on*, 2010, pp. 2070-2073.

[47] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," *Thesis (PhD). BROWN UNIVERSITY*, Source DAI-B 61/09, p. 4877, Mar 2001, 112 pages.

[48] D. Hoang and H. F. Silverman, "SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 125-128.

[49] D. Hoang and H. F. Silverman, "Stochastic particle filtering: A fast SRP-PHAT single source localization algorithm," in *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09. IEEE Workshop on*, 2009, pp. 213-216.

[50] H. Do and H. F. Silverman, "A Fast Microphone Array SRP-PHAT Source Location Implementation using Coarse-To-Fine Region Contraction(CFRC)," in *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, 2007, pp. 295-298.

[51] M. Cobos, A. Marti, and J. J. Lopez, "A Modified SRP-PHAT Functional for Robust Real-Time Sound Source Localization With Scalable Spatial Sampling," *Signal Processing Letters, IEEE,* vol. 18, pp. 71-74, 2011.

[52] P. Stoica and R. L. Moses, *Introduction to spectral analysis* vol. 57: Prentice Hall Upper Saddle River, New Jersey, 1997.

[53] H. Jianguo, X. Da, L. Xiong, and Z. Qunfei, "Maximum Likelihood DOA Estimator Based On Importance Sampling," in *TENCON 2006. 2006 IEEE Region 10 Conference*, 2006, pp. 1-4.

[54] M. Agrawal and S. Prasad, "A modified likelihood function approach to DOA estimation in the presence of unknown spatially correlated Gaussian noise using a uniform linear array," *Signal Processing, IEEE Transactions on,* vol. 48, pp. 2743-2749, 2000.

[55] P. Stoica and K. C. Sharman, "Maximum likelihood methods for direction-of-arrival estimation," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 38, pp. 1132-1143, 1990.

[56] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramer-Rao bound: further results and comparisons," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 38, pp. 2140-2150, 1990.

[57] Y. Bresler and A. Macovski, "Exact maximum likelihood parameter estimation of superimposed exponential signals in noise," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 34, pp. 1081-1089, 1986.

[58] D. Kundu, "Modified MUSIC algorithm for estimating DOA of signals," *Signal Processing*, vol. 48, pp. 85-90, 1996.

[59] D. Manolakis, V. Ingle, and S. Kogon, *Statistical and adaptive signal processing* vol. 4: McGraw-Hill Boston MA, 2000.

[60] N. Odachi, H. Shoki, and Y. Suzuki, "High-speed DOA estimation using beamspace MUSIC," in *Vehicular Technology Conference Proceedings, 2000. VTC 2000-Spring Tokyo. 2000 IEEE 51st*, 2000, pp. 1050-1054 vol.2.

[61] L. Ta-Sung, "Fast implementation of root-form eigen-based methods for detecting closely spaced sources," *Radar and Signal Processing, IEE Proceedings F*, vol. 139, pp. 288-296, 1992.

[62] R. J. Weber and H. Yikun, "Analysis for Capon and MUSIC DOA estimation algorithms," in *Antennas and Propagation Society International Symposium, 2009. APSURSI '09. IEEE*, 2009, pp. 1-4.

[63] H. Hwang, Z. Aliyazicioglu, M. Grice, and A. Yakovlev, "Direction of Arrival Estimation using a Root-MUSIC Algorithm," *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 2, 2008.

[64] J. S. Thompson, P. M. Grant, and B. Mulgrew, "Performance of spatial smoothing algorithms for correlated sources," *Signal Processing, IEEE Transactions on*, vol. 44, pp. 1040-1046, 1996.

[65] W. Peng, W. Pan-Pan, Z. Guo-jun, and X. Ji-jun, "Spatial smoothing algorithm based on acoustic vector sensor array," in *Computer Application and System Modeling (ICCASM), 2010 International Conference on*, 2010, pp. V14-27-V14-31.

[66] W. A. Gardner, "Simplification of MUSIC and ESPRIT by exploitation of cyclostationarity," *Proceedings of the IEEE*, vol. 76, pp. 845-847, 1988.

[67] T. B. Lavate, V. K. Kokate, and A. M. Sapkal, "Performance Analysis of MUSIC and ESPRIT DOA Estimation Algorithms for Adaptive Array Smart Antenna in Mobile Communication," in *Computer and Network Technology (ICCNT), 2010 Second International Conference on*, 2010, pp. 308-311.

[68] M. Shukla and R. M. Hegde, "Significance of the MUSIC-group delay spectrum in speech acquisition from distant microphones," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 2738-2741.

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

[70] Y. Jiang, P. Stoica, Z. Wang, and J. Li, "Capon beamforming in the presence of steering vector errors and coherent signals," *11th Annual Workshop on Adaptive Sensor Array Processing (ASAP 2003),* 2003.

[71] J. Li and P. Stoica, "Versatile robust Capon beamforming: theory and applications," in *Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2004,* 2004, pp. 38-42.

[72] P. Stoica, Z. Wang, and J. Li, "Robust capon beamforming," *IEEE Signal Processing letters,* June 2003, Vol. 10, Issue 6, pp. 172-175.

[73] M. Hawkes and A. Nehorai, "Acoustic vector-sensor beamforming and Capon direction estimation," *Signal Processing, IEEE Transactions on,* vol. 46, pp. 2291-2304, 1998.

[74] J. Benesty, C. Jingdong, and H. Yiteng, "A generalized MVDR spectrum," *Signal Processing Letters, IEEE,* vol. 12, pp. 827-830, 2005.

[75] E. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New Insights Into the MVDR Beamformer in Room Acoustics," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 18, pp. 158-170, 2010.

[76] B. D. van Veen, "Minimum variance beamforming with soft response constraints," *Signal Processing, IEEE Transactions on,* vol. 39, pp. 1964-1972, 1991.

[77] Y. C. Eldar, A. Nehorai, and P. S. La Rosa, "A Competitive Mean-Squared Error Approach to Beamforming," *Signal Processing, IEEE Transactions on,* vol. 55, pp. 5143-5154, 2007.

[78] D. D. Feldman and L. J. Griffiths, "A projection approach for robust adaptive beamforming," *Signal Processing, IEEE Transactions on,* vol. 42, pp. 867-876, 1994.

[79] N. L. Owsley, "Signal Subspace Based Minimum-Variance Spatial Array Processing," in *Circuits, Systems and Computers, 1985. Nineteeth Asilomar Conference on*, 1985, pp. 94-97.

[80] B. D. Carlson, "Covariance matrix estimation errors and diagonal loading in adaptive arrays," *Aerospace and Electronic Systems, IEEE Transactions on,* vol. 24, pp. 397-401, 1988.

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

[82] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *Signal Processing, IEEE Transactions on,* vol. 47, pp. 2677-2684, 1999.

[83] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *Signal Processing, IEEE Transactions on,* vol. 49, pp. 1614-1626, 2001.

[84] B. Widrow, P. E. Mantey, L. J. Griffiths, and B. B. Goode, "Adaptive antenna systems," *Proceedings of the IEEE,* vol. 55, pp. 2143-2159, 1967.

[85] P. C. Loizou, *Speech Enhancement: Theory and Practice*: CRC Press, 2007.

[86] J. P. L. a. N. Brokx, S.G., "Intonation and the perception of simultaneous voices," *Journal of Phonetics,* pp. 23-26., 1982.

[87] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*: Wiley-IEEE Press, 2006.

[88] A. Amehraye, D. Pastor, and A. Tamtaoui, "Perceptual improvement of Wiener filtering," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 2081-2084.

[89] Y. Hu and P. Loizou, "Incorporating a psychoacoustical model in frequency domain speech enhancement," *Signal Processing Letters, IEEE,* vol. 11, pp. 270-273, 2004.

[90] Y. Hu and P. C. Loizou, "A perceptually motivated approach for speech enhancement," *Speech and Audio Processing, IEEE Transactions on,* vol. 11, pp. 457-465, 2003.

[91] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, pp. 373-385, 1998.

[92] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87.*, 1987, pp. 177-180.

[93] G. Zenton, T. Kah-Chye, and B. T. G. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," *Speech and Audio Processing, IEEE Transactions on,* vol. 7, pp. 510-524, 1999.

[94] J. Benesty, *Springer Handbook of Speech Processing*: Springer, 2007.

[95] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *Speech and Audio Processing, IEEE Transactions on,* vol. 12, pp. 561-571, 2004.

[96] E. Habets, S. Gannot, I. Cohen, and P. Sommen, "Joint dereverberation and residual echo suppression of speech signals in noisy environments," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 16, pp. 1433-1451, 2008.

[97] C. Jingdong, J. Benesty, and H. Yiteng, "A Minimum Distortion Noise Reduction Algorithm With Multiple Microphones," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 16, pp. 481-493, 2008.

[98] M. S. Brandstein, "An event-based method for microphone array speech enhancement," in *Acoustics, Speech, and Signal Processing, 1999. ICASSP '99. Proceedings., 1999 IEEE International Conference on*, 1999, pp. 953-956 vol.2.

[99] S. Makino, T. W. Lee, and H. Sawada, *Blind speech separation*: Springer Verlag, 2007.

[100] A. Hyvarinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Netw,* vol. 13, pp. 411-30, May-Jun 2000.

[101] S. C. Douglas, M. Gupta, H. Sawada, and S. Makino, "Spatio&ndash;Temporal FastICA Algorithms for the Blind Separation of Convolutive Mixtures," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, pp. 1511-1520, 2007.

[102] S. C. Douglas, H. Sawada, and S. Makino, "A spatio-temporal fastICA algorithm for separating convolutive mixtures," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 2005, pp. v/165-v/168 Vol. 5.

[103] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks,* vol. 13, pp. 411-430, 2000.

[104] Y. Li, D. Powers, and J. Peach, "Comparison of blind source separation algorithms," *Advances in Neural Networks and Applications,* pp. 18–21, 2000.

[105] B. Gillespie, H. Malvar, and D. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on,* 2002, pp. 3701-3704.

[106] B. Yegnanarayana and P. Murthy, "Enhancement of reverberant speech using LP residual signal," *Speech and Audio Processing, IEEE Transactions on,* vol. 8, pp. 267-281, 2002.

[107] A. Biswas, "Advances in perceptual stereo audio coding using linear prediction techniques," *Dissertation Abstracts International,* vol. 68, ed, 2007.

[108] M. Thomas, N. Gaubitch, J. Gudnason, and P. Naylor, "A practical multichannel dereverberation algorithm using multichannel DYPSA and spatiotemporal averaging," *Applications of signal processing to audio and acoustics, IEEE workshop on*, 2007, pp. 50-53.

[109] B. Yegnanarayana, S. R. Mahadeva Prasanna, and K. Sreenivasa Rao, "Speech enhancement using excitation source information," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on,* 2002, pp. I-I.

[110] S. Gay and J. Benesty, *Acoustic signal processing for telecommunication*: Springer Netherlands, 2000.

[111] M. Goldstein, "Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener," *Speech Communication,* vol. 16, pp. 225-244, 1995.

[112] D. B. Pisoni, H. C. Nusbaum, and B. G. Greene, "Perception of synthetic speech generated by rule," *Proceedings of the IEEE,* vol. 73, pp. 1665-1676, 1985.

[113] D. Pisoni and S. Hunnicutt, "Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '80.*, 1980, pp. 572-575.

[114] ITU-T, "Methods for subjective determination of transmission quality," in *ITU-T RECOMMENDATION P.800,* ed. Helsinki: ITU-T, 1996.

[115] ITU-T, "862-perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *International Telecommunication Union-Telecommunication Standardisation Sector (ITU-T),* 2001.

[116] L. Neovius and P. Raghavendra, "Evaluation of comprehension of KTH text-to-speech with “listening speed” paradigm," *Speech Transmission Laboratory Quarterly Progress and Status Report, 2,* vol. 3, pp. 21–29, 1993.

[117] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 14, pp. 1462-1469, 2006.

[118] P. A. Naylor, N. D. Gaubitch, and E. A. P. Habets, "Signal-based performance evaluation of dereverberation algorithms," *Journal of Electrical and Computer Engineering,* vol. 2010, p. 1, 2010.

[119] R. Gray, A. Buzo, A. Gray, Jr., and Y. Matsuyama, "Distortion measures for speech processing," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 28, pp. 367-376, 1980.

[120] G. Chen, S. N. Koh, and I. Y. Soon, "Enhanced Itakura measure incorporating masking properties of human auditory system," *Signal Processing,* vol. 83, pp. 1445-1456, 2003.

[121] R. Huber and B. Kollmeier, "PEMO-Q&amp;#8212;A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 14, pp. 1902-1911, 2006.

[122] M. Karjalainen, "A new auditory model for the evaluation of sound quality of audio systems," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.*, 1985, pp. 608-611.

[123] H. Xiaopeng, H. Guiming, and Z. Xiaoping, "PEAQ-based psychoacoustic model for perceptual audio coder," in *Advanced Communication Technology, 2006. ICACT 2006. The 8th International Conference*, 2006, pp. 5 pp.-1823.

[124] D. Courville. (2007 - 2010). *Ambisonic Studio: Native B Format Recording*. Available: http://www.radio.uqam.ca/ambisonic/native_b.html

[125] E. Benjamin and Chen, "The Native B-Format Microphone," 2005.

[126] Knowles. (2005). *NR Series Microphones* [PDF]. Available: http://www.knowles.com/search/prods_pdf/NR-23158-000.pdf

[127] Knowles. (2005). *EK/EY series microphones*. Available: http://www.knowles.com/search/prods_pdf/EK-23132-000.pdf

[128] M. E. Lockwood, D. L. Jones, R. C. Bilger, C. R. Lansing, W. D. O’Brien, B. C. Wheeler, and A. S. Feng, "Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms," *The Journal of the Acoustical Society of America,* vol. 115, p. 379, 2004.

[129] A. Farina, "Advancements in impulse response measurements by sine sweeps," *Audio engineering society convention 128,* 2007, pp. 5-8.

[130] K. T. Wong and M. D. Zoltowski, "Self-initiating MUSIC-based direction finding in underwater acoustic particle velocity-field beamspace," *Oceanic Engineering, IEEE Journal of,* vol. 25, pp. 262-273, 2000.

[131] K. T. Wong and M. D. Zoltowski, "Root-MUSIC-based azimuth-elevation angle-of-arrival estimation with uniformly spaced but arbitrarily oriented velocity hydrophones," *Signal Processing, IEEE Transactions on,* vol. 47, pp. 3250-3260, 1999.

[132] M. Kawanishi, R. Maruta, N. Ikoma, H. Kawano, and H. Maeda, "Sound target tracking in 3D using particle filter with 4 microphones,"*SICE ,Annual Conference on,* 2008, pp. 1427-1430.

[133] H. Atmoko, D. Tan, G. Tian, and B. Fazenda, "Accurate sound source localization in a reverberant environment using multiple acoustic sensors," *Measurement Science and Technology,* vol. 19, p. 024003, 2008.

[134] N. Roman and D. Wang, "Binaural tracking of multiple moving sources," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 16, pp. 728-739, 2008.

[135] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments,"*Acoustic speech and signal processing (ICASSP 02), Conference on*, 2002, pp. 3021-3024.

[136] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *Speech and Audio Processing, IEEE Transactions on,* vol. 11, pp. 826-836, 2003.

[137] B. Gunel, H. Hacihabiboglu, and A. M. Kondoz, "Intensity vector direction exploitation for exhaustive blind source separation of convolutive mixtures," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 41-44.

[138] B. Gunel, H. Hacihabiboglu, and A. M. Kondoz, "Intensity vector direction exploitation for exhaustive blind source separation of convolutive mixtures," *Acoustic speech and signal processing (ICASSP 09), Conference on* 2009.

[139] IEEE Subcommittee, "IEEE Recommended Practice for Speech Quality Measurements," ed, 1969.

[140] M. Vondrasek and P. Pollak, "Methods for speech SNR estimation: Evaluation tool and analysis of VAD dependency," *Radioengineering*, vol. 14, p. 7, 2005.

[141] P. Kabal. (2007). *Voice Activity Detection (Peter Kabal ed.)* [M file]. Available: http://wwwmmsp.ece.mcgill.ca/Courses/2007-2008/ECSE412B/Project/MATLAB/VAD.m

[142] L. Anthony, B. Herbert, and K. Walter, "Multidimensional Localization of Multiple Sound Sources Using Blind Adaptive MIMO System Identification," in *Multisensor Fusion and Integration for Intelligent Systems, 2006 IEEE International Conference on*, 2006, pp. 7-12.

[143] F. Nesta, P. Svaizer, and M. Omologo, "Robust two-channel TDOA estimation for multiple speaker localization by using recursive ICA and a state coherence transform," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 4597-4600.

[144] F. Nesta and M. Omologo, "Generalized State Coherence Transform for multidimensional TDOA estimation of multiple sources," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. PP, pp. 1-1, 2011.

[145] T. Nishikawa, H. Saruwatari, and K. Shikano, "Fast-convergence blind separation of more than two sources combining ICA and beamforming," in *Nonlinear Signal and Image Processing, 2005. NSIP 2005. Abstracts. IEEE-Eurasip*, 2005, p. 17.

[146] D. Hoang and H. F. Silverman, "A method for locating multiple sources from a frame of a large-aperture microphone array data without tracking," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 301-304.

[147] G. Lathoud and M. Magimai-Doss, "A sector-based, frequency-domain approach to detection and localization of multiple speakers," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 2005, pp. iii/265-iii/268 Vol. 3.

[148] E. D. Di Claudio, R. Parisi, and G. Orlandi, "A clustering approach to multi-source localization in reverberant rooms," in *Sensor Array and Multichannel Signal Processing Workshop. 2000. Proceedings of the 2000 IEEE*, 2000, pp. 198-201.

[149] E. D. Di Claudio, R. Parisi, and G. Orlandi, "Multi-source localization in reverberant environments by ROOT-MUSIC and clustering," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, 2000, pp. II921-II924 vol.2.

[150] L. Byoung-gi and C. JongSuk, "Multi-source sound localization using the competitive k-means clustering," in *Emerging Technologies and Factory Automation (ETFA), 2010 IEEE Conference on*, 2010, pp. 1-7.

[151] J. A. Hartigan, *Clustering Algorithms*: John Wiley & Sons Inc, 1975.

[152] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR),* vol. 31, pp. 264-323, 1999.

[153] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters,* vol. 31, pp. 651-666, 2010.

[154] M. Jeub and P. Vary, "Enhancement of reverberant speech using the CELP postfilter," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 3993-3996.

[155] SPIB. *The Signal Processing Information Base (SPIB).* Available: http://spib.rice.edu/spib.html

[156] J.R.Deller, Jr., J.G.Proakis, and J.H.L.Hansen, *DISCRETE-TIME PROCESSING OF SPEECH SIGNALS*, 1993.

[157] Y. Jung-Lang and Y. Chien-Chung, "Generalized eigenspace-based beamformers," *Signal Processing, IEEE Transactions on,* vol. 43, pp. 2453-2461, 1995.

[158] P. Kroon and B. Atal, "Predictive coding of speech using analysis-by-synthesis techniques," *Signals, Systems and Computers, 1990 Conference Record Twenty-Fourth Asilomar Conference on* , vol.2, no., pp.664, 5-7 Nov 1990.

[159] D. Campbell, "The ROOMSIM user guide (V3. 3)," ed.

[160] S. M. Schimmel, M. F. Muller, and N. Dillier, "A fast and accurate "shoebox" room acoustics simulator," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 241-244.

[161] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *Neural Networks, IEEE Transactions on,* vol. 10, pp. 626-634, 1999.

[162] M. Pedersen, D. Wang, J. Larsen, and U. Kjems, "Two-microphone separation of speech mixtures," *Neural Networks, IEEE Transactions on*, vol. 19, pp. 475-492, 2008.

[163] M. S. Pedersen and C. M. Nielsen, "Gradient flow convolutive blind source separation," in *Machine Learning for Signal Processing, 2004. Proceedings of the 2004 14th IEEE Signal Processing Society Workshop*, 2004, pp. 335-344.

[164] *Soundfield: An Introduction*. Available: www.soundfield.com/feature.htm.

[165] S. Rickard and F. Dietrich, "DOA estimation of many W-disjoint orthogonal sources from two mixtures using DUET," in *Statistical Signal and Array Processing, 2000. Proceedings of the Tenth IEEE Workshop on*, 2000, pp. 311-314.

[166] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 516-527, 2011.

[167] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing,* vol. 87, pp. 1833-1847, 2007.

[168] C. Weiping, Z. Xiaoyan, and W. Zhenyang, "Localization of Multiple Speech Sources Based on Sub-band Steered Response Power," in *Electrical and Control Engineering (ICECE), 2010 International Conference on*, 2010, pp. 1246-1249.

[169] S. Cao, L. Li, and X. Wu, "Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise," *The Journal of the Acoustical Society of America*, vol. 129, p. 2227, 2011.

[170] J. S. Garofolo, *Darpa Timit: Acoustic-phonetic Continuous Speech Corps CD-ROM*: US Dept. of Commerce, National Institute of Standards and Technology, 1993.

[171] A. K. Swain and W. Abdulla, "Estimation of LPC parameters of speech signals in noisy environment," in *TENCON 2004. 2004 IEEE Region 10 Conference*, 2004, pp. 139-142 Vol. 1.

[172] A. Trabelsi, F. R. Boyer, Y. Savaria, and M. Boukadoum, "Improving LPC analysis of speech in additive noise," in *Circuits and Systems, 2007. NEWCAS 2007. IEEE Northeast Workshop on*, 2007, pp. 93-96.

[173] N. Gaubitch, D. Ward, and P. Naylor, "Statistical analysis of the autoregressive modeling of reverberant speech," *The Journal of the Acoustical Society of America,* vol. 120, p. 4031, 2006.

[174] L. Zhao and M. W. Hoffman, "Application of microphone array for speech coding in noisy environment," in *Signals, Systems and Computers, 1996. 1996 Conference Record of the Thirtieth Asilomar Conference on*, 1996, pp. 45-49 vol.1.

[175] Core-Sound. (1990). *Core Sound TetraMic*. Available: http://www.core-sound.com/TetraMic/1.php

[176] D. T. Hemingson and M. J. Sarisky, "Improvements on a Low-Cost Experimental Tetrahedral Ambisonic Microphone,"*Audio Engineering Society Convention 128*" May 2010