Speech, Audio, Image and Video Technology Laboratory
School of Engineering Systems

SPEECH RECOGNITION USING AD-HOC MICROPHONE ARRAYS

Ivan Himawan
B.Eng(Hons)

SUBMITTED AS A REQUIREMENT OF
THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT
QUEENSLAND UNIVERSITY OF TECHNOLOGY
BRISBANE, QUEENSLAND
APRIL 2010
Keywords

ad-hoc array, microphone array, beamforming, array calibration, blind speech separation, speech enhancement, automatic speech recognition
Abstract

While close talking microphones give the best signal quality and produce the highest accuracy from current Automatic Speech Recognition (ASR) systems, the speech signal enhanced by microphone array has been shown to be an effective alternative in a noisy environment. The use of microphone arrays in contrast to close talking microphones alleviates the feeling of discomfort and distraction to the user. For this reason, microphone arrays are popular and have been used in a wide range of applications such as teleconferencing, hearing aids, speaker tracking, and as the front-end to speech recognition systems.

With advances in sensor and sensor network technology, there is considerable potential for applications that employ ad-hoc networks of microphone-equipped devices collaboratively as a virtual microphone array. By allowing such devices to be distributed throughout the users’ environment, the microphone positions are no longer constrained to traditional fixed geometrical arrangements. This flexibility in the means of data acquisition allows different audio scenes to be captured to give a complete picture of the working environment. In such ad-hoc deployment of microphone sensors, however, the lack of information about the location of devices and active speakers poses technical challenges for array signal processing algorithms which must be addressed to allow deployment in real-world applications. While not an ad-hoc sensor network, conditions approaching this have in effect been imposed in recent National Institute of Standards and Technology (NIST) ASR evaluations on distant microphone recordings of meetings. The NIST evaluation data comes from multiple sites, each with different and often loosely specified distant microphone configurations.

This research investigates how microphone array methods can be applied for ad-hoc microphone arrays. A particular focus is on devising methods that are robust to unknown microphone placements in order to improve the overall speech
quality and recognition performance provided by the beamforming algorithms. In ad-hoc situations, microphone positions and likely source locations are not known and beamforming must be achieved blindly. There are two general approaches that can be employed to blindly estimate the steering vector for beamforming. The first is direct estimation without regard to the microphone and source locations. An alternative approach is instead to first determine the unknown microphone positions through array calibration methods and then to use the traditional geometrical formulation for the steering vector. Following these two major approaches investigated in this thesis, a novel clustered approach which includes clustering the microphones and selecting the clusters based on their proximity to the speaker is proposed. Novel experiments are conducted to demonstrate that the proposed method to automatically select clusters of microphones (ie, a sub-array), closely located both to each other and to the desired speech source, may in fact provide a more robust speech enhancement and recognition than the full array could.
# Contents

Keywords i

Abstract iii

List of Tables x

List of Figures xiv

Acronyms & Abbreviations xix

Authorship xxii

Acknowledgements xxiii

1 Introduction 1

1.1 Motivation and Overview 1

1.2 Aims and Objectives 3

1.3 Scope of Thesis 4

1.4 Outline of Thesis 5

1.5 Original Contributions of Thesis 7

1.6 Publications Resulting from Research 8

1.6.1 International Journal Publications 8

1.6.2 International Conference Publications 8

2 Microphone Array Processing Techniques 11

2.1 Introduction 11

2.2 Sound Wave Propagation 11

2.3 Apertures 13
2.3.1 Continuous Apertures .............................................. 13
2.3.2 Linear Apertures Theory ........................................ 14
2.3.3 Linear Sensor Array .............................................. 16
2.3.4 Spatial Aliasing ................................................... 18
2.4 Microphone Array Beamforming Techniques ................. 19
2.5 Noise Fields ......................................................... 22
2.5.1 Spatially Uncorrelated Noise ................................. 22
2.5.2 Spherical Isotropic Noise ...................................... 23
2.5.3 Cylindrically Isotropic Noise ................................ 23
2.6 Evaluation of Beamformers ......................................... 24
2.6.1 Spatial Directivity Pattern .................................... 24
2.6.2 Array Gain and White Noise Gain ............................ 24
2.6.3 Sensitivity to Array Imperfections ........................... 26
2.6.4 Directivity and Directivity factor ............................ 27
2.7 Delay-sum Beamforming ............................................ 27
2.8 Superdirective Beamforming ....................................... 28
2.9 Adaptive Array Processing ........................................ 29
2.10 Post Filtering ......................................................... 31
2.11 Summary ............................................................ 33

3 A Review of Speech Recognition using Microphone Arrays and Research Directions 35
3.1 Introduction .......................................................... 35
3.2 Fundamentals of Speech Recognition .............................. 35
3.2.1 Speech Feature Extraction .................................... 36
3.2.2 Decoding .......................................................... 38
3.2.3 Parameter Estimation .......................................... 41
3.2.4 Model Adaptation .............................................. 41
3.2.5 Measuring Speech Recognition Performance ............ 42
3.3 Microphone Array for Robust Speech Recognition .......... 43
3.4 Research Directions .................................................. 45
4 Automatic Array Calibration 53

4.1 Introduction 53

4.2 Array Shape Calibration 54

4.3 Microphone Array Shape Calibration in Diffuse Noise Fields 58
  4.3.1 Multidimensional Scaling 59
  4.3.2 Experimental Results 60

4.4 Speaker Localisation 64
  4.4.1 SRP-PHAT Localisation 67

4.5 Automatic Calibration 68
  4.5.1 Experimental Results 69

4.6 Beamforming Approach to Blind Speech Separation 71
  4.6.1 System Overview 72
  4.6.2 Experimental Results 74
  4.6.3 Speech Recognition Experiments 81

4.7 Summary 83

5 Blind Beamforming 85

5.1 Introduction 85

5.2 Blind Estimation Approach 86
  5.2.1 Summary of Blind Beamforming Approaches 88

5.3 Experimental Results 89
  5.3.1 Database Specifications 89
  5.3.2 Baseline Experiments 91
  5.3.3 Comparison of Blind Approaches 93

5.4 Discussion 96

5.5 Summary 98
6 Clustered Approach for Ad-hoc Microphone Arrays

6.1 Introduction .................................................. 101
6.2 Theoretical Justification ................................. 103
6.3 Inter-Microphone Proximity Measure ................. 105
  6.3.1 Rule-based Clustering ................................. 106
  6.3.2 Spectral Clustering .................................. 108
6.4 Ranking Clusters ........................................... 109
6.5 Clustered Beamforming ................................... 111
  6.5.1 Closest Cluster Beamforming ....................... 111
  6.5.2 Weighted Cluster Combination Beamforming .... 111
6.6 Experimental Setup ....................................... 112
  6.6.1 Data Set A ............................................. 112
  6.6.2 Data Set B ............................................. 113
  6.6.3 Data Set C ............................................. 114
  6.6.4 Data Set D ............................................. 115
6.7 Experimental Results ..................................... 115
  6.7.1 Microphone Clustering Evaluation ................. 115
  6.7.2 Cluster Proximity Ranking ......................... 117
  6.7.3 Blind Beamforming Performance Evaluation .... 119
  6.7.4 Experiments on Weighted Cluster Combination ... 123
6.8 Discussion .................................................... 126
  6.8.1 Clustering of Microphones ......................... 126
  6.8.2 Blind Beamforming Experiments ................. 126
  6.8.3 Speech Recognition Experiments ............... 129
  6.8.4 Weighted Cluster Combination Experiments .... 130
6.9 Conclusion .................................................. 131

7 Conclusions and Future Research .................. 135

7.1 Summary of Contributions ................................. 135
7.2 Additional Contributions .................................. 138
7.3 Future Research ............................................ 139

Bibliography .................................................... 141
## List of Tables

3.1 Speech recognition performance of robust beamforming techniques using ground-truth microphone positions. .......................... 50

4.1 Flying distances (in km) between Australian cities. .................. 61

4.2 Accuracy of inter-microphone distance estimates of the calibrated array (in cm). ................................................................. 64

4.3 Accuracy of microphone location estimates of the calibrated array (in cm). ................................................................. 64

4.4 Accuracy of inter-microphone distance estimates of the calibrated circular array from MONC recording (in cm). ...................... 70

4.5 Accuracy of microphone location estimates of the calibrated circular array from MONC recording (in cm). ...................... 70

4.6 Accuracy of speaker localisation results in terms of azimuth (degrees), elevation (degrees), and radius (metres) error using the ground truth array geometry (Ground truth), calibrated array from MONC (Calibrated 1), and calibrated array from Figure 4.3(c). (Calibrated 2). ................................................................. 70

4.7 Mean, standard deviation, minimum and maximum inter-microphone distance error. All measurements are in centimeters. ............ 75

4.8 Results measuring accuracy of positions estimates in terms of the mean, standard deviation, minimum and maximum distance error. All measurements are in centimetres. ................................. 75

4.9 % Accuracy of speaker localisation results in the development set using actual microphone positions, in terms of azimuth (degrees), elevation (degrees), and radius (metres) error. ......................... 77
4.10 % Accuracy of speaker localisation results in the development set using estimated microphone positions, in terms of azimuth (degrees), elevation (degrees), and radius (metres) error. 77

4.11 % Accuracy of the merged speaker localisation results in the development set using actual and estimated microphone positions, in terms of azimuth (degrees), elevation (degrees), and radius (metres) error. 78

4.12 % word error rate on development data set of the proposed system. 82

4.13 % word error rate on evaluation data set of the proposed system. 82

5.1 Design of beamforming techniques 89

5.2 % word error rate for headset, lapel, and single distant microphone 90

5.3 % word error rate for blind delay-sum beamforming 92

5.4 % word error rates for stationary speaker task using a circular array in various speaker positions. 93

5.5 % word error rates for stationary speaker task using a circular array in various speaker positions. 93

5.6 % word error rates for moving speaker task using a circular array in various speaker movements. 94

5.7 % word error rates for stationary speaker task using an ad-hoc array in various speaker positions. 94

5.8 % word error rates for moving speaker task using ad-hoc array in various speaker movements. 96

6.1 Clustering results in terms of average $acp$, $asp$ and $K$ over 20 ad-hoc geometries. 117

6.2 Microphone membership and the average proximity score $\bar{D}_k$ (in ms) to clusters for the ad-hoc array setup in Data Sets B, C and D. 117

6.3 Segmental SNR (dB) of ad-hoc array cluster beamforming in Data Set B. Results are averaged over 30 utterances. The best result for each scenario is highlighted. 120
6.4 PESQ measures of ad-hoc array cluster beamforming in Data Set B. Results are averaged over 30 utterances. The best result for each scenario is highlighted. .................................................. 121

6.5 Segmental SNR (dB) of ad-hoc array cluster beamforming in data set C for closest cluster and second closest cluster. Results are averaged over 30 utterances. The best result for each scenario is highlighted. .................................................. 121

6.6 PESQ measures of ad-hoc array cluster beamforming in data set C for closest cluster and second closest cluster. Results are averaged over 30 utterances. The best result for each scenario is highlighted. 122

6.7 % word error rate from various blind beamforming methods on Data Set D. For comparison, WER of 26.1% is achieved using the original clean corpus recordings on the same ASR system. . . . . 123

6.8 % word error rate of weighted cluster delay-sum combination. . 124
List of Figures

2.1 Spherical coordinates of a point in space. . . . . . . . . . . . . . . 13
2.2 Incoming signal received by a linear aperture. . . . . . . . . . . . 15
2.3 Directivity pattern of a linear aperture. . . . . . . . . . . . . . . . 16
2.4 Directivity pattern for 7-elements linear sensor array with 0.1m
    spacing as a function of angle $\phi$ and frequency $f$. . . . . . . . 17
2.5 Example of spatial aliasing for a linear sensor array with four ele-
    ments. Note the appearance of grating lobes for $d = \lambda_{\text{min}}$ (right)
    due to spatial aliasing. . . . . . . . . . . . . . . . . . . . . . . . . . . 18
2.6 Signal model received by the microphone array. . . . . . . . . . . 19
2.7 Directivity pattern of uniformly weighted linear array. . . . . . . 25
2.8 Directivity pattern of uniformly weighted linear array in polar plot. 26
2.9 Generalised sidelobe canceller structure. . . . . . . . . . . . . . . 30
2.10 Robust generalised sidelobe canceller structure. . . . . . . . . . 31
2.11 Beamformer with post-filter. . . . . . . . . . . . . . . . . . . . . . 32

3.1 Typical speech recognition system, outlining both the training of
    speech models and the testing process. . . . . . . . . . . . . . . . . 36
3.2 Mel-frequency filter bank for N=8. . . . . . . . . . . . . . . . . . 37
3.3 An example of 5-state HMM with left-to-right topology. . . . . . 39
3.4 Conventional microphone array based speech recognition system. . 43
3.5 Meeting room configuration in MONC recording . . . . . . . . . . 50
3.6 Word error rate versus microphone placement error for various
    robust beamforming approaches. . . . . . . . . . . . . . . . . . . . 51
4.1 Measured coherence function of microphone pair with a distance of 0.2m. The theoretical sinc function is plotted for true microphone distances. .................................................. 59

4.2 Map of Australia based on flying distances of major cities, using classical multidimensional scaling. .............................................................. 61

4.3 Actual (circles) and its corresponding estimated microphone location (pluses) for linear array (top), square array (middle), and circular array (bottom). .................................................. 63

4.4 Spatial directivity patterns of calibrated arrays for delay-sum beamformer (top) and superdirective beamformer (bottom) with weights calculated using true microphone positions (dotted lines) and estimated positions (straight lines). In each case, the beamformer is directed towards a speaker at 1 m, and response is plotted at 1 kHz. 65

4.5 System diagram of microphone array beamforming technique for blind source separation in PASCAL SSC2 challenge. ......................... 72

4.6 The actual (circles) and estimated (pluses) microphone positions for sub-array1 and sub-array2. .................................................. 76

4.7 The spectrogram of the original overlapped speech, enhanced superdirective and post-filter output, and the headset microphone recording of the female speech. The corresponding binary mask for the post-filter is also plotted. .................................................. 79

4.8 The spectrogram of the original overlapped speech, enhanced superdirective and post-filter output, and the headset microphone recording of the male speech. The corresponding binary mask for the post-filter is also plotted. .................................................. 80

5.1 Edinburgh meeting room, according to [41] (measurements in cm). 91

5.2 Directivity patterns plots. .................................................. 95
6.1 Box plot of the TDOA values applied to 8-element microphone arrays. The lines are drawn for the lower quartile, median, and upper quartile values. Whiskers extend to one standard deviation above and below the mean (circle) of the data. The expected ground truth of TDOA values are shown in asterisk. In the x axis, TDOA_{21} corresponds to the TDOA between the second closest and the closest microphone in the array, similarly, TDOA_{31} corresponds to the TDOA between the third closest and the closest microphone in the array, and so on.

6.2 Theoretical magnitude squared coherence as a function of distance between two microphones. The sampling frequency is 16 kHz.

6.3 Meeting room used in Data Sets B and D (measurements in cm).

6.4 Meeting room used in Data Set C (measurements in cm).

6.5 Cluster assignments on the ground truth positions for four geometries, with one geometry per row. In each case, (A) shows the sub-arrays obtained from known microphone positions while (B) and (C) shows the result of rule-based clustering and spectral clustering from the measured noise coherence. The centre microphone in a cluster is shown by encircling line for rule-based clustering.

6.6 The delay-sum SNR improvement of weighted cluster combination for noise condition N1 in different speaker positions by varying weight \( \alpha_1 \) for the closest cluster and weight \( \alpha_2 \) for the second closest cluster.

6.7 Box plot of the TDOA values applied to closest cluster, second closest cluster, and all microphones for speaker at position S1 with noise condition N1 from Data Set B. The attributes of plots are similar to box plot in Figure 6.1.
# Acronyms & Abbreviations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMI</td>
<td>Augmented multi-party interaction</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial neural network</td>
</tr>
<tr>
<td>ASR</td>
<td>Automatic speech recognition</td>
</tr>
<tr>
<td>BSS</td>
<td>Blind source separation</td>
</tr>
<tr>
<td>CHIL</td>
<td>Computers in the human interaction loop</td>
</tr>
<tr>
<td>CSTR</td>
<td>Centre for speech technology research</td>
</tr>
<tr>
<td>DOA</td>
<td>Direction of arrival</td>
</tr>
<tr>
<td>GCC</td>
<td>Generalised cross correlation</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian mixture model</td>
</tr>
<tr>
<td>GSC</td>
<td>Generalised sidelobe canceller</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden markov model</td>
</tr>
<tr>
<td>ICSI</td>
<td>International computer science institute</td>
</tr>
<tr>
<td>LCMV</td>
<td>Linear constrained minimum variance</td>
</tr>
<tr>
<td>LIMABEAM</td>
<td>Likelihood maximising beamformer</td>
</tr>
<tr>
<td>LMS</td>
<td>Least mean square</td>
</tr>
<tr>
<td>MC-WSJ-AV</td>
<td>Multi-channel wall street journal audio visual (database)</td>
</tr>
<tr>
<td>MDM</td>
<td>Multiple distant microphone</td>
</tr>
<tr>
<td>MDS</td>
<td>Multidimensional scaling</td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel-frequency cepstral coefficients</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum likelihood</td>
</tr>
<tr>
<td>MLLR</td>
<td>Maximum likelihood linear regression</td>
</tr>
<tr>
<td>MONC</td>
<td>Multichannel overlapping number corpus (database)</td>
</tr>
<tr>
<td>MSC</td>
<td>Magnitude squared coherence</td>
</tr>
<tr>
<td>MUSIC</td>
<td>Multiple signal classification</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>MVDR</td>
<td>Minimum variance distortionless response</td>
</tr>
<tr>
<td>NIST</td>
<td>National institute of standards and technology</td>
</tr>
<tr>
<td>PDA</td>
<td>Personal digital assistant</td>
</tr>
<tr>
<td>PHAT</td>
<td>Phase transform</td>
</tr>
<tr>
<td>PLP</td>
<td>Perceptual linear predictive</td>
</tr>
<tr>
<td>PMC</td>
<td>Parallel model combination</td>
</tr>
<tr>
<td>RGSC</td>
<td>Robust generalised sidelobe canceller</td>
</tr>
<tr>
<td>ROVER</td>
<td>Recogniser output voting error reduction</td>
</tr>
<tr>
<td>SCOT</td>
<td>Smoothed coherence transform</td>
</tr>
<tr>
<td>SDM</td>
<td>Single distant microphone</td>
</tr>
<tr>
<td>SINR</td>
<td>Signal-to-interference-plus-noise ratio</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal to noise ratio</td>
</tr>
<tr>
<td>SRP-PHA T</td>
<td>Steered response power with phase transform</td>
</tr>
<tr>
<td>SVD</td>
<td>Singular value decomposition</td>
</tr>
<tr>
<td>TDOA</td>
<td>Time difference of arrival</td>
</tr>
<tr>
<td>UEDIN</td>
<td>University of Edinburgh</td>
</tr>
<tr>
<td>WER</td>
<td>Word error rate</td>
</tr>
<tr>
<td>WNG</td>
<td>White noise gain</td>
</tr>
<tr>
<td>WSJ</td>
<td>Wall street journal (database)</td>
</tr>
</tbody>
</table>
Authorship

The work contained in this thesis has not been previously submitted for a degree or diploma at any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signed: ______________________

Date: ______________________
Acknowledgements

It is not possible to thank everyone who crosses my path during the course of my PhD. First and most importantly, I would like to thank my parents, brother and sisters who have been very supportive during years of my PhD research. Their patience and understanding enable me to go through the darkest days with laughter. Special mention to my fiancée, Elsa Feiliany, who always be there when I need support and encouragement, be patient with me with her advice and wisdom. Thank you for your constant prayers to strengthen my faith in this journey of love.

I would also like to thank my principal supervisor, Professor Sridha Sridharan for his guidance and encouragement throughout my PhD program. His encouragement and financial support has given me the opportunity to discover the world of research. Special mention to my associate supervisor, Dr. Iain McCowan for his constant guidance and support throughout my course of study. This dissertation may not be completed if not because his unwavering belief in my ability to do research.

During my PhD research, I was fortunate to visit the Centre for Speech Technology Research (CSTR) at Edinburgh, Scotland. I would like to thank Professor Steve Renals and Dr. Mike Lincoln for giving me the opportunity to work with them as Augmented Multi-party Interaction (AMI) trainee. This was an invaluable experience, which gave me the opportunity to do research at an outstanding institution. My full appreciation goes to Dong Wang, Dr. Joe Frankel, and Dr. Simon King for your collaboration which I hope will continue in the future. To Chao Qin whom I met during my internship, thank you for your friendship. I wish you all well for your future research endeavours.

The members of the SAIVT laboratory must also be acknowledged, for the
great atmosphere they created as well as their generosity for helping colleagues in needs. I would like to mention (in alphabetical order) Brendan Baker, Chris McCool, Clinton Fookes, David Dean, Eddie Wong, Frank Lin, Mark Cox, Michael Mason, Mitchell McLaren, Patrick Lucey, Robbie Vogt, Roy Wallace, Ruwan Lakemond, Simon Denman, Terry Martin, and Tristan Kleinschmidt. I wish them all every success in life whatever path they may take.

Finally, thanks to all my friends here in Brisbane especially members of the Indonesian Catholic Charismatic Prayer group and St. Stephen’s Cathedral Choir, and to my PhD colleagues Leonard and Arnold Wiliem, for their ongoing support and assistance during the course of my PhD.
Chapter 1

Introduction

1.1 Motivation and Overview

Automatic speech recognition technology has in recent years attracted significant attention as a means of providing robust human-to-computer interaction. State-of-the-art ASR systems have recently attained a sufficiently high performance to be deployed in practical applications. One of the examples of ASR in action, is a piece of commercial software called “Dragon Naturally Speaking”. This software enables a user to speak naturally with computer in order to dictate e-mail or word documents and also to edit them in real time. To achieve such high performance of recognition accuracy, the user is recommended to wear a close-talking microphone (e.g. headset microphone) supplied by the software.

An alternative to a close-talking microphone is a microphone array which consists of multiple microphones arranged in a purposeful geometry. The spatial configuration of the array enables the desired speech signal to be filtered from interfering signals such as competing speech or room reverberation since they are usually originated from different points in space. This process of focusing attention to a particular direction is called beamforming [82]. Therefore, a microphone array could be located at a significant distance from the speaker whilst still providing a hands-free signal acquisition.

Microphone arrays have significant advantages over close talking microphones since they enable natural interaction of the speaker due to their unobtrusive
nature. Since it alleviates distraction and the feeling of discomfort after a long use, the speaker could interact naturally with computer or between other speakers in multi-speaker environment such as in meetings. While close-talking microphones give the highest accuracy from the current ASR systems, speech signal enhanced by microphone array has been shown to be an effective alternative in a noisy environment.

With advances in sensor and sensor network technology, an ad-hoc network of microphone-equipped devices can be arranged as a virtual microphone array [57]. By allowing flexibility in the method of data acquisition compared to the traditional microphone array arrangements, it allows a more pervasive or “real-world” technology to develop, which would be a major benefit to many applications. While not an ad-hoc sensor network, one example of data collected using arrays with loosely specified distant microphone configurations is that used in the NIST rich transcription meeting recognition evaluation. The data contain recordings of natural meeting interactions captured from a number of sites, each with their own microphone configurations. These configurations are unknown to the evaluation participants who must use the recorded data to perform speech recognition of the meetings [21].

Unfortunately most traditional beamforming algorithms are based on assumptions that the array is stationary and has a known geometry. This requires prior knowledge of microphone characteristics as well as its precise location - none of which can be guaranteed for an ad-hoc microphones deployment. Furthermore, in order to achieve adequate performance, microphone array processing algorithms need to be able to minimise environmental noise and interference. Therefore, a number of aspects in ad-hoc microphone array processing such as unknown position, gain, and synchronisation, need to be considered. These factors which previously have not been considered in the conventional algorithms are now seen to be affecting the performance of speech recognition.

The main motivation of this thesis is to conduct research into practical aspects of ad-hoc microphone arrays which will enable the future deployment of ad-hoc microphone array ASR systems. This is particularly applicable in the situation when both microphone and speaker locations are unknown. This is important
because one of the key components of any practical microphone array acquisition systems is the ability to robustly localise and track the active speaker. Ad-hoc microphone arrays processing have a great potential in practical applications of speech recognition due to their ability to handle multiple distant microphones distributed throughout the user’s environment while providing robustness to noise and reverberation.

1.2 Aims and Objectives

Ad-hoc microphone array processing for speech recognition is an emerging and very broad topic of research, with only a minimal amount of work having been conducted in the area to date. As a result, the existing research on the topic has several major limitations which this thesis aims to address. This dissertation aims to build on the existing body of research by extending the current approaches to enable practical deployment of ad-hoc microphone arrays for speech recognition applications, and also to discover new methods by which enhancement systems can be optimised for improved speech recognition performance.

The general aims of this thesis are:

(i) To investigate the impact of using ad-hoc microphone arrays on speech recognition performance.

(ii) To propose novel approaches to beamforming from ad-hoc microphone arrays when microphone positions and likely source locations are not known.

(iii) To consider the implementation of the proposed algorithms within the constraints of meeting room environments.

More specifically, the research objectives are:

(i) To review and identify limitations in the current state of research into ad-hoc microphone arrays for speech recognition applications.

(ii) To quantify the word accuracy performance of ASR system when there is error in microphone placements from the original assumed positions.
(iii) To propose two general approaches to blindly estimate the steering vector for beamforming when microphone positions and likely source locations are not known.

(iv) To propose a novel approach to blind speech separation using microphone array in multi-speaker environments.

(v) To propose novel techniques in dealing with randomly placed microphones and speaker positions.

(vi) To assess each of the proposed techniques and report their performance based on speech recognition accuracy using recorded data from the meeting room environments.

1.3 Scope of Thesis

Microphone array speech recognition is in itself a very broad area of research, being the combination of the field of array processing and speech recognition. For ad-hoc deployment of microphones, there are many practical aspects which need to be considered so that they can be used to perform effective speech enhancement and recognition. In order to give this thesis a tight focus in reaching its aims and objectives, it is therefore very important to limit the scope of the research to the following fields and applications:

Microphone array beamforming techniques for speech enhancement

Beamforming techniques exploit the array positions to provide spatial discrimination in order to direct the beam to the desired speaker location. The techniques are popular for speech enhancement since the principle has been known for a long time and therefore widely used for the front-end of speech recognition systems. Alternative techniques such as multi-channel blind signal separation and subspaced-based algorithms are usually more computationally expensive and often introduce considerable distortion to the desired speech. As a result these are not widely accepted in the speech recognition community.
Unknown microphone and speaker positions In ad-hoc situations, array geometry and speaker positions are mostly unknown. The focus of this thesis is to devise methods or new methodologies to perform speech enhancement without prior knowledge of locations by using only audio signals collected by the microphones. Apart from unknown locations, the problem of ad-hoc deployment of microphone arrays may consist of dealing with variation in microphone responses and issues in synchronisation. To tackle all of these issues simultaneously would require a huge amount of time and resources and, as such, the latter two problems are not dealt within this thesis and are considered as directions for further research.

Medium-sized vocabulary speech recognition Medium-size task offers an intermediate task between simple digit recognition and large vocabulary conversational speech recognition. The task gives a better indication of the performing system in the practical applications while limiting the complexity of the ASR system.

Meeting room environment This thesis focuses mainly on the meeting task that occurs in the typical office meeting rooms. All experiments were recorded in this environment. Experiments were conducted in three scenarios which are common in real meetings: a single seated active speaker, a moving speaker, and overlapping speech from concurrent speakers. While in the real meetings, there is no limit to the number of meeting participants, the study in this thesis has been confined to the specific cases of one and two active speakers around a meeting table.

1.4 Outline of Thesis

The remainder of this thesis is organised as follows:

Chapter 2 presents the fundamental theory of array processing and key microphone array beamforming techniques for the purpose of speech enhancement.
Chapter 3 reviews the theory behind speech recognition systems along with a literature review of microphone-array-based speech recognition. Current limitations on microphone array front-end processing for speech recognition in meetings are identified, and from these, research directions are proposed. The chapter concludes with a comparative study of robust beamforming techniques when there is error in microphone placements from the original assumed positions.

Chapter 4 presents a review of array calibration techniques. This leads to the novel approach of estimating the microphone positions from the unknown array geometry. One of the applications of this technique is a novel approach to blind speech separation using microphone array.

Chapter 5 investigates two general approaches to blindly estimate the steering vector for beamforming when microphone positions and likely source locations are not known.

Chapter 6 proposes a novel clustered approach to blind beamforming from ad-hoc microphone arrays. Without using any prior geometrical information, microphones are first grouped into localised clusters, which are then ranked according to their relative distance from a speaker. Beamforming is then performed using either the closest microphone cluster, or a weighted combination of clusters. The clustered algorithms are compared to the full set of microphones in experiments on a database recorded on different ad-hoc array geometries. These experiments evaluate the methods in terms of signal enhancement as well as the performance on a medium vocabulary speech recognition task.

Chapter 7 summarises the work contained in this thesis, highlighting major research findings. Avenues for future work and development are also discussed.
1.5 Original Contributions of Thesis

In this thesis a number of original contributions were made to the field of microphone array and ASR in general. These are summarised as:

(i) *Comparative study of various beamforming techniques when there is error in microphone placements from the original assumed positions in terms of speech recognition accuracy.*

(ii) *Two approaches to beamforming from unknown microphone and speaker locations.* In contrast, the conventional beamforming requires prior knowledge of array geometry and speaker locations. The first approach is through an automatic calibration technique in which a steering vector for beamforming could be derived and the second is the direct estimation of a steering vector from the signals. While a direct estimation technique has been used previously for beamforming from ad-hoc arrays, the work in this thesis compares both approaches in terms of speech recognition accuracy.

(iii) *Analysis of a novel array shape calibration technique in diffuse noise field [47].* The novel array calibration technique differs from other methods in the literature in which it does not require purpose-built devices or known calibration signals, but relies simply on the background noise signal.

(iv) *Blind separation of overlapped speech using microphone array beamforming techniques.* The proposed technique does not require a priori knowledge of the microphone and speaker locations commonly assumed, making it comparable to the typical blind source separation approaches.

(v) *A novel clustered approach for robust blind beamforming from ad-hoc arrays.* The proposed approach includes novel blind clustering of microphones using magnitude squared coherence, cluster proximity ranking based on Time Difference of Arrival (TDOA), and two different methods of beamforming using these clusters, termed closest cluster beamforming and weighted cluster combination beamforming.
(vi) *Experiments on real ad-hoc microphone geometry in the meeting room environments.* Comparison between delay-sum, MVDR, and superdirective beamforming in ad-hoc array scenarios throughout the thesis.

In addition, a novel posterior approach to microphone-array-based speech recognition is proposed and presented in the conference publication number (iii) in the following section. Also, the application of FPGA-implemented delay-sum beamforming for in-car speech enhancement and recognition is presented in the publication number (ii). As these works are not directly related to the principle topic of this thesis, speech recognition using ad-hoc microphone arrays, these works are not included in the body of the thesis. Nevertheless they are still within the broad topic of microphone-array-based speech recognition and the relevant publications are listed.

### 1.6 Publications Resulting from Research

The following fully-refereed publications have been produced as a result of the work in this thesis:

#### 1.6.1 International Journal Publications


#### 1.6.2 International Conference Publications


Chapter 2

Microphone Array Processing Techniques

2.1 Introduction

This chapter is intended to give an overview of microphone array processing techniques. The most common approach to processing microphone array signals is beamforming which is a method of spatial filtering. The first part of this chapter will focus on the fundamental theory of array processing where most of the materials are summarised from [49]. The chapter then focuses on presenting the theory of a number of key microphone array beamforming techniques with particular emphasis on optimisation for different noise fields.

2.2 Sound Wave Propagation

An active speaker produces sound waves which propagate through the air until they reach each of the microphones in the room. The sound waves propagate as longitudinal waves through fluids. The movement of the molecules in the fluid in the direction of propagation produces regions of compression and expansion. Using Newton’s equations of motion considering an ideal fluid where there is no viscosity, the linear wave equation can be expressed as [49, 92]

\[ \nabla^2 x(t, p) = \frac{1}{c^2} \delta^2 x(t, p) = 0 \tag{2.1} \]
where \( x(t, p) \) is a function representing the sound pressure of an acoustic wave field at time instant \( t \) for a point in space at Cartesian coordinates \( p = [x, y, z]^T \). The \( \nabla^2 \) is the Laplacian operator and \([.]^T\) denotes the transpose. \( c \) which is the speed of propagation, depends upon the atmospheric pressure and temperature for a given specific wave type and medium. For the acoustic waves in air at the sea level, the \( c \) is approximately 340\( \text{ms}^{-1} \).

The acoustic wave propagates from the sound source as monochromatic spherical waves, where the amplitude is decaying at a rate proportional to the distance from the source. When this distance is large, they then can be considered as plane waves to simplify the mathematical analysis. For the plane wave, the solution to the differential wave equation is given as \[ x(t, p) = A(t) \exp[j(\omega t - k \cdot p)] \quad (2.2) \]

where \( A(t) \) is the wave amplitude and \( \omega = 2\pi f \) is the frequency in radians per second. The \( k \) is the wavenumber vector, which indicates the speed and direction of wave propagation is given by

\[
k = \frac{2\pi}{\lambda} \begin{bmatrix} \sin \theta \cos \phi & \sin \theta \sin \phi & \cos \theta \end{bmatrix}
\quad (2.3)
\]

where \( \lambda \) is the wavelength related to \( c \) by \( \lambda = c/f \). The \( \theta \) and \( \phi \) represent the waveform position in space in spherical coordinates as illustrated in Figure 2.1 for elevation and azimuth respectively.

Another solution to the differential wave equation is given by spherical wave solved as \[ x(t, p) = -\frac{A(t)}{4\pi p} \exp[j(\omega t - kp)] \quad (2.4) \]

where \( p = |p| \) is the Euclidian distance from the source and \( k \) is the scalar wavenumber given by \( 2\pi/\lambda \). The spherical wave solution shows that the signal amplitude decreases at a proportional rate to the distance of observation. In general, when the point of observation is close to the emitting source as in near field condition, the wavefront of the propagating wave is modelled as spherical waves as opposed to the plane waves for the far field.

The solution of the wave equation in Equation 2.1 shows that the propagating acoustic signals can be expressed both in time and space. This relation implies
that one can reconstruct the original signal from the signal which is sampled in
time and space as long as it satisfies the Nyquist rule. The time sampling is done
to acquire the digital signal while spatial sampling is done by the microphone
array.

2.3 Apertures

An aperture is defined as an element that transmits or receives radiation. An
acoustic aperture is an electroacoustic transducer that receives propagating waves
and converts them into electrical currents (microphone), or vice versa (loud-
speaker). In this context, microphone array can be considered as acoustic trans-
ducers, placed in different locations as a spatially sampled version of continuous
aperture. By means of sampling theory, the fundamental theory of sensor arrays
can be described from the continuous aperture principles.

2.3.1 Continuous Apertures

The continuous aperture is given as a finite area that is responsive to the imping-
ing signal. When the incoming signal arrives at the receiving aperture, the amount
of signal seen is dependent on the spatial orientation of aperture as illustrated in

Figure 2.1: Spherical coordinates of a point in space.
Figure 2.2. In this context, the *aperture function* is defined as a response to the incoming signal. The response of the receiving element with impulse response $a(t, p)$ to the propagating wave $x(\tau, p)$ is given as the result of convolution:[92]

$$x_R(t, p) = \int_{-\infty}^{\infty} x(\tau, p)a(t-\tau, p) d\tau$$

(2.5)

with the frequency domain equivalent by means of Fourier transform

$$X_R(f, p) = X(f, p)A(f, p)$$

(2.6)

The response of the receiving aperture depends on the amount of signal seen by the aperture as the signal’s direction of arrival varies. When the response is evaluated for all arrival directions, the overall aperture response can be seen from its *directivity pattern* or *beampattern* as a function of frequency and direction of arrival. The far-field directivity pattern of the aperture function $A_R$ is given by

$$D_R(f, \alpha) = \mathcal{F}_p\{A_R(f, p)\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} A_R(f, p)e^{j2\pi \alpha \cdot p} dp$$

(2.7)

where $\mathcal{F}_p(.)$ is the three dimensional Fourier transform, with $p = [x_a, y_a, z_a]^T$ is the spatial location of a point along the aperture and

$$\alpha = \frac{1}{\lambda}\begin{bmatrix} \sin \theta \cos \phi \\ \sin \theta \sin \phi \\ \cos \theta \end{bmatrix}$$

(2.8)

indicates the direction of wave.

### 2.3.2 Linear Apertures Theory

In order to investigate some properties of the aperture directivity pattern oriented towards array processing, it is useful to simplify the above equations by considering linear aperture. For a linear shaped aperture with length $L$ located along the x-axis, the directivity pattern simplifies to

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

(2.9)

where

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

(2.10)

For the case of linear aperture, considering:
2.3. Apertures

1. Aperture function is constant over all frequencies.

2. The signal is assumed to arrive from the far-field, thus

$$|p| > \frac{2L^2}{\lambda}$$  \hspace{1cm} (2.11)

The aperture function may be written as

$$A_R(x_a) = \text{rect}(x_a/L)$$  \hspace{1cm} (2.12)

where $\text{rect}(.)$ is the rectangular function. In this case, the resulting directivity pattern simplifies to

$$D_R(f, \alpha_x) = \mathcal{F}\{\text{rect}(x_a/L)\} = L\text{sinc}(\alpha_x L)$$  \hspace{1cm} (2.13)

where

$$\text{sinc}(x) \equiv \frac{\sin(x)}{x}$$  \hspace{1cm} (2.14)

The plot of directivity pattern corresponding to a uniform aperture function is shown in Figure 2.3. The directivity pattern covers the area of $-\lambda/L \leq \alpha_x \leq \lambda/L$ is referred to as the main lobe and its extent is termed the beam width. From the pattern plot, the zeros are located at $\alpha_x = m\lambda/L$. Note that the beam width of
a linear aperture is given by $2\lambda/L$, or in terms of frequency $2c/fL$. This shows that the beam width is inversely proportional to the product $fL$, hence for the fixed aperture length, the beam width will decrease with increasing frequency.

### 2.3.3 Linear Sensor Array

As mentioned above, a sensor array could be considered as a sampled version of continuous aperture. In this case, the aperture is only excited at a finite number of discrete positions. The overall response of the array can be determined as the superposition of all individual element responses.

For an array with $N$ elements, where each element has a different complex frequency response $e_n(f, x)$, the aperture function of the array is now written as

$$A(f, x_a) = \sum_{n=-N^{-1}}^{N^{-1}} w_n(f)e_n(f, x_a - x_n) \quad (2.15)$$

where $w_n(f)$ is the complex weighting for element $n$ and $x_n$ is the spatial position of such an element on the $x$-axis. The far-field directivity pattern from discrete array aperture function can be computed by substituting Equation 2.15 into Equation 2.9 to obtain

$$D(f, \alpha_x) = \sum_{n=-N^{-1}}^{N^{-1}} w_n(f)E_n(f, \alpha_x)e^{j2\pi\alpha x_n} \quad (2.16)$$
2.3. Apertures

Figure 2.4: Directivity pattern for 7-elements linear sensor array with 0.1m spacing as a function of angle $\phi$ and frequency $f$.

where $E_n(f, \alpha_x)$ is the directivity pattern of element $n$. If all elements have an identical frequency response (that is $E_n(f, \alpha_x) = E(f, \alpha_x), \forall n$), the directivity pattern can be simplified to

$$D(f, \alpha_x) = \sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} w_n(f)e^{j2\pi \alpha_x x_n}$$  \hspace{2cm} (2.17)

Considering only a horizontal directivity pattern of line array aligned to the x-axis where all elements are equally spaced by $d$ metres, the directivity pattern becomes

$$D(f, \alpha_x) = \sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} w_n(f)e^{j\frac{2\pi L}{d} \cos \phi}$$ \hspace{2cm} (2.18)

The term $w_n(f)$ is the complex weighting function which can be expressed in terms of its magnitude $a_n(f)$ and phase component $\varphi_n(f)$ as

$$w_n(f) = a_n(f)e^{j\varphi_n(f)}$$ \hspace{2cm} (2.19)

Modifying the weight will enable the full control of the shape and the angular location of beamformer’s directivity pattern. In Section 2.4, the method to determine this weight will be derived which in turn leads to various beamformer
2.3.4 Spatial Aliasing

![Diagram](image)

Figure 2.5: Example of spatial aliasing for a linear sensor array with four elements. Note the appearance of grating lobes for $d = \lambda_{\text{min}}$ (right) due to spatial aliasing.

The minimum sampling frequency required to avoid aliasing in order to sample the continuous time-signal is given by a Nyquist frequency. In an analogous manner, the sensor array implements spatial sampling in which aliasing could occur by the appearance of grating lobes in the directivity pattern when the signal is under-sampling in the spatial domain.

To avoid aliasing, the temporal sampling theorem states that the signal must be sampled at a rate $f_s$

$$f_s = \frac{1}{T_s} \geq 2f_{\text{max}} \quad (2.20)$$

where $f_{\text{max}}$ is the maximum frequency in the signal’s frequency spectrum. In a similar fashion, to avoid spatial aliasing it requires that

$$f_{x_s} = \frac{1}{d} \geq 2f_{x_{\text{max}}} \quad (2.21)$$

where $f_{x_s}$ is the spatial sampling frequency in samples per metre and $f_{x_{\text{max}}}$ is the highest spatial frequency component in the spectrum of the signal. The $f_{x_{\text{max}}}$ is inversely proportional to $\lambda_{\text{min}}$ which is the minimum wavelength in the signal of interest. This leads to the requirement that

$$d < \frac{\lambda_{\text{min}}}{2} \quad (2.22)$$
The polar plot of the horizontal directivity pattern to illustrate the effect of spatial aliasing is shown in Figure 2.5.

\section*{2.4 Microphone Array Beamforming Techniques}

In practice, the sensors could be positioned arbitrarily at any point in space. Randomly placed sensors can work collaboratively as a microphone array to spatially sample a sound wave impinging on them.

Consider a desired signal received by an omni-directional microphone $i$ sampled at discrete time $t$ as illustrated in Figure 2.6. The microphone output is an attenuated and delayed version of the desired signal $a_is(t-\tau_i)$ and noise $v_i$ given by

$$x_i(t) = a_is(t-\tau_i) + v_i(t)$$ (2.23)

where the signal and noise are assumed to be statistically independent. In the frequency domain, the signal model is written as

$$x_i(f) = a_is(f)e^{-j2\pi f \tau_i} + v_i(f)$$ (2.24)

The array signal model of $N$ microphones is stacked into a vector and written
as

\[ x(f) = s(f)d(f) + v(f) \]  \hspace{1cm} (2.25)

where \( x(f) \) represents frequency domain coefficients at discrete frequency \( f \). The data vector \( x(f) \) defined as:

\[ x(f) = [x_1(f) \cdots x_i(f) \cdots x_N(f)]^T \]  \hspace{1cm} (2.26)

and

\[ v(f) = [v_1(f) \cdots v_i(f) \cdots v_N(f)]^T \]  \hspace{1cm} (2.27)

The \( d(f) \) represents the array steering vector which depends on the actual microphone and source location. In the near field, \( d(f) \) is given by [7]:

\[ d(f) = [a_1e^{-j2\pi f\tau_1} \cdots a_ie^{-j2\pi f\tau_i} \cdots a_Ne^{-j2\pi f\tau_N}]^T, \]  \hspace{1cm} (2.28)

where gain scaling of microphone \( a_i \) is

\[ a_i = \frac{d_{ref}}{d_i} \]  \hspace{1cm} (2.29)

and delay \( \tau_i \) is

\[ \tau_i = \frac{d_i - d_{ref}}{c} \]  \hspace{1cm} (2.30)

The \( d_i \) and \( d_{ref} \) denote the Euclidian distance between the source and the microphone \( i \), or the reference microphone, respectively. The \( d_i \) and \( d_{ref} \) is given by:

\[ d_i = ||q - p_i|| \]  \hspace{1cm} (2.31)

\[ d_{ref} = ||q - p_{ref}|| \]  \hspace{1cm} (2.32)

where \( q \), \( p_i \), and \( p_{ref} \) are the the spatial locations of source \( q \), microphone \( p_i \), and microphone \( p_{ref} \) in the Cartesian coordinates respectively.

In order to retrieve the desired signal \( s(f) \), \( x(f) \) which represents the received signals from all sensors is processed by a frequency domain filter weights \( w(f) \). The weight vector \( w(f) \) is defined as:

\[ w(f) = [w_1(f) \cdots w_i(f) \cdots w_N(f)]^T \]  \hspace{1cm} (2.33)

The beamformer output is the sum of \( N \) weighted microphone outputs given by

\[ y(f) = \sum_{i=1}^{N} w_i^H(f)x_i(f) \]  \hspace{1cm} (2.34)
2.4. Microphone Array Beamforming Techniques

or, in the matrix form (omitting the frequency dependence index \( f \) without loss of generality)

\[
y = w^H x
\]

(2.35)

where operator \((\cdot)^H\) denotes Hermitian transpose. The inverse Fourier transform results in time domain output signal \( y(t) \).

To find the beamformer weights using distortionless criterion, it requires that in the absence of noise, the beamformer output \( y(f) = s(f) \). Thus, the variance of the output signal \( y(f) \) needs to be minimised in the presence of noise. For \( y(f) = s(f) + y_n(f) \), the mean square of the output noise

\[
E[|y_n|^2] = w^H \Phi_{vv} w
\]

(2.36)

needs to be minimised. The \( \Phi_{vv} \) is the cross power spectral density matrix of noise defined as

\[
\Phi_{vv} = \begin{pmatrix}
\Phi_{v_1 v_1} & \Phi_{v_1 v_2} & \cdots & \Phi_{v_1 v_N} \\
\Phi_{v_2 v_1} & \Phi_{v_2 v_2} & \cdots & \Phi_{v_2 v_N} \\
\vdots & \vdots & \ddots & \vdots \\
\Phi_{v_N v_1} & \Phi_{v_N v_2} & \cdots & \Phi_{v_N v_N}
\end{pmatrix}
\]

(2.37)

where \( \Phi_{v_1 v_1}(f) \) and \( \Phi_{v_1 v_2}(f) \) are auto- and cross-power spectral densities, respectively.

The no distortion implies that \( w^H d = 1 \), which signifies an undistorted signal response in the desired look direction. Therefore, the constrained minimisation problem which has to be solved is:

\[
\min_w w^H \Phi_{vv} w
\]

(2.38)

subject to distortionless constraint given by:

\[
w^H d = 1
\]

(2.39)

The solution to this problem can be solved by using the method of Lagrange multiplier to find the minimum of a function subject to constraints [31, 81]. The well known solution is usually termed the Minimum Variance Distortionless Response (MVDR) weights, given by:
Chapter 2. Microphone Array Processing Techniques

\[ w_{MV,DR} = \Phi^{-1}_{vv} \frac{d}{d^H \Phi^{-1}_{vv} d} \]  

(2.40)

Assuming that the noise field is stationary and spatially homogeneous, the noise cross power spectral density matrix can be expressed in terms of the coherence matrix \( \Gamma_{vv} \):

\[
\Gamma_{vv} = \begin{pmatrix}
\Gamma_{v1v1} & \Gamma_{v1v2} & \cdots & \Gamma_{v1vN} \\
\Gamma_{v2v1} & \Gamma_{v2v2} & \cdots & \Gamma_{v2vN} \\
\vdots & \vdots & \ddots & \vdots \\
\Gamma_{vNv1} & \Gamma_{vNv2} & \cdots & \Gamma_{vNvN}
\end{pmatrix}
\]  

(2.41)

where \( \Gamma_{v_iv_j} \) is given by:

\[
\Gamma_{v_iv_j}(f) \triangleq \frac{\Phi_{v_iv_j}(f)}{\sqrt{\Phi_{v_iv_i}(f) \Phi_{v_jv_j}(f)}}
\]  

(2.42)

Thus, Equation 2.40 can be expressed in terms of coherence function as:

\[
w_{MV,DR} = \frac{\Gamma_{vv}^{-1} d}{d^H \Gamma_{vv}^{-1} d}
\]  

(2.43)

In the next section, the coherence matrix which represent the spatial characteristics of noise is presented in detail. The knowledge of the noise field the beamformer operates on is important in order design a beamformer with maximum performance.

2.5 Noise Fields

Coherence is a frequency domain measurement of correlation. Earlier work published by Cook et al [12] measures the spatial correlation of two omni-directional microphones in an isotropic noise field. Coherence function is useful in the design and analysis of the performance of beamformers in the actual noise fields.

2.5.1 Spatially Uncorrelated Noise

In spatially uncorrelated noise or an incoherent noise, the correlation of noise signals received by two microphones at any given spatial location is zero. Thus,
coherence is zero in all frequencies between two different spatial locations, that is $\Gamma_{vv} = I$ (where $I$ is the identity matrix). An ideal type of this noise would be thermal noise which is generated by the equilibrium fluctuations of the electric current inside an electrical conductor and considered as spatially white. In the case of microphone arrays however, ideal incoherent noise is difficult to achieve and merely theoretical.

### 2.5.2 Spherical Isotropic Noise

In spherical isotropic noise or a diffuse noise, two spatially separated microphones receive equal energy and random phase noise signals from all directions simultaneously. This type of noise appears in several practical reverberant environments, such as inside offices or cars. For a diffuse noise field, the coherence function can be modelled as [12]:

$$\Gamma_{v_i v_j}^{\text{diff}} = \text{sinc} \left( \frac{2\pi f d_{ij}}{c} \right)$$  \hspace{1cm} (2.44)

where $d_{ij}$ is the distance between sensor $i$ and $j$, and the sinc function has been defined in Equation 2.14.

### 2.5.3 Cylindrically Isotropic Noise

An alternative to the spherical isotropic noise, a cylindrically isotropic noise field is commonly assumed to model some room acoustic fields. The cylindrically isotropic noise function is modelled by [16]:

$$\Gamma_{v_i v_j}^{\text{cyl}} = J_0 \left( \frac{2\pi f d_{ij}}{c} \right)$$  \hspace{1cm} (2.45)

where $J_0$ denotes the 0th order Bessel function of the first kind. This function suits to model babble noise as encountered in the cocktail party problem. Thus, it is usually applied in the design of speech enhancement for hearing aid applications [19, 62].
Chapter 2. Microphone Array Processing Techniques

2.6 Evaluation of Beamformers

This section presents measures that are typically used to analyse the array performance. Each of the various measures attempts to quantify the response of the array from the incoming signal in the environment the beamformer operates on.

2.6.1 Spatial Directivity Pattern

The spatial directivity pattern or beampattern of beamformer computes the response of the array to a wavefront emitted from specific direction with particular frequency. For illustrating the basic characteristics of sensor array, the linear array of uniformly weighted beamformer located along the x-axis is introduced. The spatial response for an incoming signal considering only a horizontal directivity pattern over angle $\phi$ is defined as:

$$H(f, \phi) = w^H(f)d(f, \phi)$$ (2.46)

Figure 2.7 depicts the directivity pattern of uniformly weighted 7-elements linear array with spacing between elements $d = \lambda/2$. From the beampattern plot, relative side-lobe and peak-to-zero distance are defined

1. relative side-lobe: the relative height of the first side-lobe with respect to the main-lobe.

2. peak-to-zero distance: the region between the maximum of the main lobe and the first null.

Another common method to visualise the directivity pattern is to plot $H(f, \phi)$ as a function of $\phi$ while $f$ is held fixed in polar coordinates as shown in Figure 2.6.1 for the same array example.

2.6.2 Array Gain and White Noise Gain

Array gain is defined as the Signal-to-Noise Ratio (SNR) improvement between one channel sensor signal and the output of the microphone array. From this
2.6. Evaluation of Beamformers

Figure 2.7: Directivity pattern of uniformly weighted linear array.

The SNR of one sensor signal is given by the ratio between signal power spectral density $\Phi_{ss}$ and the average noise power $\Phi_{vv}$. Therefore, the array gain is

$$ G = \frac{|w^H d|^2}{w^H \Gamma_{vv} w} $$ (2.51)

One of the important array gain measure is the White Noise Gain (WNG), which measures the ability of the array to suppress spatially uncorrelated noise.
Figure 2.8: Directivity pattern of uniformly weighted linear array in polar plot.

Substituting the coherence matrix of spatially uncorrelated noise (i.e. identity matrix $I$) into $\Gamma_{vv}$ in Equation 2.51 gives

\[ G_{WN} = \frac{|w^Hd|^2}{w^Hw} \]  

(2.52)

### 2.6.3 Sensitivity to Array Imperfections

The sensitivity of the array to position, gain, and phase errors has been shown to be inversely proportional to the white noise gain [81]. The effect of random deviations from the ideal model is the attenuation of beampattern and the decrease of white noise gain as the sensitivity increases. The sensitivity function for any array geometry is defined as

\[ T_{se} = G_{WN}^{-1} = ||w||^2 \]  

(2.53)

as long as $|w^Hd|^2 = 1$. 
2.6.4 Directivity and Directivity factor

Directivity is defined as the ratio of power pattern (squared magnitude of the beampattern) in the look direction to the power pattern averaged over all directions of incoming signals for a given frequency $f$.

$$D(f) = \frac{|H(f, \theta_0, \phi_0)|^2}{\frac{\pi}{4\pi} \int_0^{2\pi} \int_0^\pi |H(f, \theta, \phi)|^2 \sin \theta \, d\phi \, d\theta}$$

Equation 2.54 can be expressed in the matrix notation as

$$D(f) = \frac{|w^H(f)d(f)|^2}{w^H(f) \left( \frac{1}{4\pi} \int_0^{2\pi} \int_0^\pi d(f) d^H(f) \sin \theta \, d\phi \, d\theta \right) w(f)}$$

The denominator which represents the noise power at the array output due to the isotropic noise can be written as a matrix $\Gamma$ where

$$\Gamma = \frac{1}{4\pi} \int_0^{2\pi} \int_0^\pi d(f) d^H(f) \sin \theta \, d\phi \, d\theta$$

in which directivity can be concisely expressed as (by omitting the frequency index)

$$D = \frac{|w^Hd|^2}{w^H\Gamma w}$$

From this definition, the directivity factor which measures the ability of the array to suppress the spherically isotropic noise (i.e. diffuse noise) can be simply written as

$$G_{DF} = \frac{|w^Hd|^2}{w^H\Gamma_{vv} w}$$

2.7 Delay-sum Beamforming

Delay-sum beamforming is the simplest of all beamforming techniques. Delay-sum compensates for the signal delay to each microphone output appropriately. After summing outputs, the desired signal will be reinforced, while the noise signals will be effectively reduced through destructive interference. To align each channel in the common time reference, a microphone location is usually selected as a reference point. The delay for each channel $i$ is then calculated from Equation 2.30.
In the frequency domain, the signal delay is obtained by applying a phase shift to each channel signal spectrum. To normalise the output after the summation so that equal amplitude between output and input signal is obtained, the signal in each channel is scaled by a uniform gain factor that is inversely proportional to the number of microphones. The channel weight for delay-sum is:

\[ w_i(f) = \frac{1}{N} e^{-2j\pi f i} \quad (2.59) \]

An identical weight is obtained by assuming a spatially uncorrelated noise in the coherence function of MVDR solution by substituting the coherence matrix for identity matrix, \( \Gamma_{vv} = I \) in Equation 2.43. This relation shows that delay-sum beamformer is basically an optimal beamformer in a spatially uncorrelated noise field.

### 2.8 Superdirective Beamforming

Superdirective beamformer aims to maximise the array gain in a diffuse noise field (i.e. maximise the directivity factor):

\[
\max_w \left| \frac{w^H d}{w^H \Gamma_{vv}^\text{diff} w} \right|^2
\]

subject to distortionless constraint \( w^H d = 1 \). The maximisation of Equation 2.60 is equal to the minimisation of the noise power in the denominator of this equation. Therefore, the solution is basically MVDR beamformer weights in which the coherence matrix is obtained from the diffuse noise field model.

It is well known that superdirective filter weights amplify the uncorrelated noise at the lower frequency range and are sensitive to deviations from the mismatch in the assumed microphone characteristics (gain, phase, position). To prevent the excessive amplification of noise in superdirective beamformers, the commonly used technique is to impose the white noise gain constraint \([7, 13, 14]\):

\[
\left| \frac{w^H d}{w^H w} \right|^2 \geq \delta^2
\]

(2.61)

Note that, by placing the constraint in Equation 2.61, this is equal to putting constraint in the sensitivity function (i.e. limiting the norm of the filter \( ||w||^2 = w^H w \leq \delta^2 \)) \([81]\).
Incorporating the WNG constraint in the maximisation and using the method of Lagrange multipliers, the solution to this optimisation problem is to add a scalar value \( \mu \) to the main diagonal of coherence matrix:

\[
\mathbf{w}_{sd,qc} = \frac{(\Gamma_{vv}^{\text{diff}} + \mu \mathbf{I})^{-1}\mathbf{d}}{d^H(\Gamma_{vv}^{\text{diff}} + \mu \mathbf{I})^{-1}d}
\]  

(2.62)

The amount of \( \mu \) can vary from zero which results in an unconstrained superdirective to infinity which results in delay-sum. The simplest strategy to find for this value is by trial and error, the \( \mu \) value is chosen in order to get a reasonable directivity factor and white noise gain. Another strategy to find the desired \( \mu \) value which satisfy the WNG constraint is to employ a binary search between specified minimum and maximum values of \( \mu \). This results in frequency variant \( \mu \) restricted to operating frequencies of beamformer.

### 2.9 Adaptive Array Processing

Delay-sum and superdirective beamforming are both classified as fixed beamformer or data-independent, since their array processing parameters do not change dynamically over time. In the moving speaker scenario, even though the delay may change when it tracks the speaker, both methods are still considered to be fixed algorithms. In contrast, adaptive or data-dependent algorithms change their parameters to adjust to noise or speaker conditions on either a sample by sample or frame by frame basis.

Presumably, a well known adaptive beamforming is the Frost algorithm [23], which implements constrained Least Mean Square (LMS) in order to maintain the chosen frequency response in the look direction while minimising the noise output power. An alternative beamforming structure for realising the constrained LMS of Frost algorithm is the Generalised Sidelobe Canceller (GSC) [28]. In relation to Linear Constrained Minimum Variance (LCMV) beamformers in which Frost’s algorithm belongs to, GSC is basically an efficient realisation of LCMV that transform the constrained optimisation problem into an unconstrained one. Note that MVDR is being the specific case of LCMV, in which only the unity look-direction constraint is in place.
GSC consists of two main structures given in Figure 2.9, a fixed beamformer in the upper part is non-adaptive and an adaptive portion in the lower part which contains a blocking matrix and multiple input canceller.

The blocking matrix realises the directional constraint by cancelling signals from the desired direction, thus generating noise reference signals. The simplest implementation of the blocking matrix is the delay-and-subtract beamformer. The noise reference signals are then led to a multiple input canceller which is unconstrained adaptive filters that aim to cancel the remaining noise from the output of the fixed beamformer in the upper path.

In the real world speech applications, deviations to the acoustic model are commonly occurs because of the difference in microphone characteristics from previously assumed model. These errors can cause portions of the desired signals to leak into the blocking matrix. This enables the adaptive portion of GSC structure to partly reconstruct desired signal components and to substract this from the fixed beamformer output. The outcome is the target signal cancellation in the form of the attenuation of the desired signal.

To combat this problem, the Robust Generalised Sidelobe Canceller (RGSC) replaces the fixed blocking matrix in the standard GSC with an adaptive one [33] as depicted in Figure 2.10. The new structure enables the blocking matrix to adaptively track the desired signal input as speaker moves with little distortion by alleviating the phase error that occured due to the difference in the actual and the assumed steering direction. Thus, the RGSC is more robust to target
2.10 Post Filtering

To further improve the SNR of the enhanced signal after beamforming, it is a common practice to apply a post-filter at the output of the beamformer. The

signal cancellation then the standard GSC using a fixed blocking matrix. In a practical implementation of the RGSC, adaptation control is needed to prevent target signal cancellation. This is due to the fact that the desired signal for the adaptation algorithm in the blocking matrix is in the contrary to that in the multiple input canceller. Since the fixed beamformer can not output the desired signal free from the noise, the blocking matrix should only be adapted during the high SNR portion of the input signal (e.g. noisy speech period) to prevent the suppression of noise by the blocking matrix. In the multiple input canceller however, the desired signal is noise. Because the blocking matrix can not estimate the noise replica that is completely free of the desired signal, the multiple input canceller should be adapted when SNR is low (e.g. noise only period). The efficient implementation of the RGSC using a block frequency domain method is presented in [32].

While showing some benefits over a fixed beamformer, the problem of signal cancellation due to signal reflections in reverberant environments prevent these adaptive techniques to gain popularity for speech recognition applications.

Figure 2.10: Robust generalised sidelobe canceller structure.
Chapter 2. Microphone Array Processing Techniques

Multi-channel Wiener filter provides an optimal solution in the minimum mean squared error sense to the problem of noise reduction for broadband inputs. The optimal weights can be shown as the product of a real-valued scalar vector $h_{\text{post}}$ with MVDR beamformer weight [75]

$$w_{\text{opt}} = h_{\text{post}} \frac{\Phi_{vv}^{-1} d}{d^H \Phi_{vv}^{-1} d}$$ (2.63)

and $h_{\text{post}}$ is

$$h_{\text{post}} = \frac{\Phi_{ss}}{\Phi_{ss} + \Phi_{vv}}$$ (2.64)

where $\Phi_{ss}$ and $\Phi_{vv}$ are respectively the signal and noise auto spectral density at the output of the beamformer. The above equation suggests that the optimum post filter is basically a single channel Wiener filter that operates on beamformer output.

A thorough investigation of microphone arrays with post-filtering has been examined by Marro et al. [45]. In this study, the optimal interaction between the array and post-filter is achieved in the structure shown in Figure 2.11. The common problem in the formulation of post-filter transfer function $h_{\text{post}}$ is the estimation of the signal and noise auto spectral densities. A possible solution is to take advantage of multi-channel signal acquisition by estimating signal correlation between different microphone channels.

The noise reduction system proposed by Zelinski provides the first published...
solution of post-filtering for multi-microphone signal acquisition [90]. The solution which is based on a Wiener filter uses auto- and cross-spectral densities measure calculated between the microphone channels in order to control the attenuation factor of the output signal’s amplitude. The basic assumption behind this post-filter is that the noise signal between the microphones are uncorrelated, corresponding to perfectly incoherent noise. This assumption is only valid if the microphone distance is large enough and the noise field is diffuse.

In practice, for the closely spaced microphones, the noise is highly correlated in the lower frequency spectrum. To suppress the high-correlated noise, two-path processing is usually employed in which the higher frequency region follows the standard post-filter formulation while the lower frequency region uses single channel noise suppression techniques [40, 51]. The McCowan post-filter [46] on the other hand generalises the Zelinksi solution with the formulation of post-filter based on diffuse noise field coherence. It has been shown to perform better over Zelinksi post-filter in a reverberant environment in terms of speech enhancement and recognition.

2.11 Summary

In this chapter, a number of key microphone array processing techniques have been discussed. The theory and the governing equations behind beamforming techniques were described. These beamforming techniques can be characterised by the noise fields for which they have been optimised. It was shown that delay-sum is optimal in an incoherent noise field and similarly superdirective beamforming for optimal noise reduction in a diffuse noise field. An overview of adaptive beamforming techniques and the methods of post-filtering to further improve the performance of beamformers were also presented.
Chapter 3

A Review of Speech Recognition using Microphone Arrays and Research Directions

3.1 Introduction

Automatic speech recognition systems have been widely used for many applications in order to provide transcripts in video-conferencing, teleworking, and hands-free computing. For this purpose, one of the goals for noise reduction algorithm is to improve the speech recognition accuracy. In this thesis, microphone arrays are used for signal acquisition and enhancement. While SNR and mean opinion score have been generally used as a measure of speech quality, it is desirable that the proposed techniques are able to improve the overall speech recognition accuracy. The significant gain in the recognition accuracy provides a more relevant performance measure in many applications that involve the ASR systems.

3.2 Fundamentals of Speech Recognition

The speech recognition system is essentially a pattern recognition device [56]. The input to the system is a speech signal and the speech recogniser decodes it as
text. Figure 3.1 below shows the main operations of the speech recognition system. The speech signal is first parameterised by extracting relevant information for the purpose of classification, and the decoding of this information is achieved by the means of acoustic and language models to determine the most likely spoken word sequence. To recognise unknown speech, models must first be trained from the training dataset which includes the training data and the corresponding transcriptions. The trained speech models can then be used to evaluate the system’s performance using the testing dataset. In the following sub-sections, the theory and the techniques of the ASR system are described.

3.2.1 Speech Feature Extraction

The input speech signal is first transformed into a sequence of feature vectors suited for speech modelling by minimising the redundancy in speech waveform through a series of signal processing steps in the feature extraction stage. There are two main approaches to speech parameterisation. The first method is based on a smoothed power spectrum by linear predictive analysis to model the speech production using a series of filter coefficients called linear predictive coefficients. The second method is based on a filter bank analysis which tries to emulate human auditory perception. Among the features which are based on the filter bank analysis widely used for speech recognition are the Mel-Frequency Cepstral Coefficients (MFCCs) [17].
3.2. Fundamentals of Speech Recognition

To compute the MFCCs vectors, a number of pre-processing steps are applied to the raw input speech signals. The input signal is first divided into successive segments of overlapping frames. As speech is a time-varying signal, a short frame is desirable as it can be considered to be quasi-stationary. Typical values for frame size are between 10 to 25ms with a frame shift between 5 to 10ms. For the purpose of compensating for the spectral tilt which occurred during speech production, each frame is filtered with a two-tap pre-emphasis filter with a transfer function of

\[ H(z) = 1 - az^{-1} \]  \hspace{1cm} (3.1)

where \( a \) is the constant that is typically in the range 0.95 \( \leq a \leq 0.98 \).

Following the pre-emphasis operation, a window is applied to each speech frame in order to prevent distortion in the frequency domain due to the effect of the Gibbs Phenomenon [36]. A commonly used window is a Hamming window defined by

\[ w(n) = 0.54 - 0.46 \cos \left( \frac{2\pi n}{L} \right), 0 \leq n \leq L \]  \hspace{1cm} (3.2)

Each windowed frame is then transformed into the frequency domain using a short time Fourier transform. The triangular-shaped filters for Mel-frequency analysis, uniformly spaced in the Mel-frequency scale which is defined by

\[ Mel(f) = 2595 \log_{10} \left( 1 + \frac{f}{700} \right) \]  \hspace{1cm} (3.3)

as shown in Figure 3.2 are applied to the magnitude spectrum of the speech signal. The purpose of using the Mel-scale is to mimic the auditory processing of...
human ear since empirical evidence shows that the human ear resolves frequencies in a non-linear manner.

The energies for each filter-bank beneath each triangular window is then calculated by weighting the spectral magnitude values by its corresponding filter gain and the results accumulated. The logarithm of the total energy in each band is computed and the resulting log energy coefficients are finally decorrelated using the discrete cosine transform from which is truncated according to the order of MFCCs analysis. The signal energy which is simply log energy of the input frame, and the time derivatives of the first order and the second order of the speech features are referred to as delta and delta² parameters respectively and are often appended to the feature space in order to have a more robust parameter set [24].

3.2.2 Decoding

The vast majority of speech recognition systems employ a Hidden Markov Model (HMM) as a statistical model to characterise and recognise the temporal evolution of speech features. In these systems, a distinct HMM is assigned to model acoustic units such as phonemes, groups of phonemes, or words. The HMM can be characterised by

1. a finite number of states

2. the state transition probabilities which govern the probability of moving to the next state

3. the output probability density function associated to each state

The example of HMM with 5 number of states (entry, 3 emitting states, and exit state) is shown in Figure 3.3. In this example of HMM with left-to-right topology, the only permitted transitions are to return back to the current state or to move the next state to the right. The transitions between states is governed by a transition matrix, \( \mathbf{A} = \{a_{ij}\} \). The element of this matrix, \( a_{ij} \) defines the probability of moving from state \( i, s_i \) at current time \( t \) to state \( j, s_j \) at time \( t + 1 \). The state probability densities which describe the distribution of observation
Figure 3.3: An example of 5-state HMM with left-to-right topology.

Vectors belonging to a particular state is usually modelled by a mixture of M Gaussians distributions. Thus, the output probability for observation vectors at time $t$, $o_t$ belonging to state $j$ is represented as

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm} N(o_t; \mu_{jm}, \Sigma_{jm})$$

$$= \sum_{m=1}^{M} c_{jm} \frac{1}{\sqrt{(2\pi)^{n}|\Sigma_{jm}|}} \exp \left\{ -\frac{1}{2} (o_t - \mu_{jm})^T \Sigma_{jm}^{-1} (o_t - \mu_{jm}) \right\} \quad (3.4)$$

where $c_{jm}$ is the mixture weight of the $m^{th}$ multi-variate Gaussian with a mean vector $\mu_{jm}$ and covariance matrix $\Sigma_{jm}$ of state $j$ of the HMM. Given the state transition matrix, $A$, the set of probability densities $B = \{b_j(.)\}$ for all states, and the initial state occupancy probability, $P(q_1 = s_j|\lambda) = \pi_j$, the HMM is completely defined by the set of parameter $\lambda = (A, B, \pi)$.

In the HMM design, the model comprises the states and transition probabilities and the underlying state sequence is said to be hidden. Because the underlying state sequence $Q$ is unknown, the total probability to generate a sequence of observation feature vectors $O$:

$$O = \{o_1, o_2, \cdots, o_T\} \quad (3.5)$$

given a specific HMM must be evaluated for every possible state sequence $Q = [q_1, q_2, \cdots, q_T], 1 \leq q_t \leq S$. This can be expressed as

$$P(O|\lambda) = \sum_{all\,Q} P(O, Q|\lambda) = \sum_{all\,Q} P(O|Q, \lambda)P(Q|\lambda) \quad (3.6)$$
Assuming statistical independence of observations,

\[ P(O|Q, \lambda) = \prod_{t=1}^{T} p(o_t|q_t, \lambda) = b_{q_1}(o_1)b_{q_2}(o_2) \cdots b_{q_T}(o_T) \]  

(3.7)

and the probability of state sequence \( Q \), is given by

\[ P(Q|\lambda) = \pi_{q_1}a_{q_1q_2}a_{q_2q_3} \cdots a_{q_{T-1}q_T} \]  

(3.8)

where \( \pi_{q_1} \) is the probability of the HMM initially in state \( q_1 \). Substituting equations 3.7 and 3.8 into 3.6,

\[ P(O|\lambda) = \sum_{q_1, q_2, \cdots, q_T} \pi_{q_1}b_{q_1}(o_1)a_{q_1q_2}b_{q_2}(o_2) \cdots a_{q_{T-1}q_T}b_{q_T}(o_T) \]  

(3.9)

Clearly, the solution to this problem is mathematically intractable since it requires the summing over all possible state sequences to compute the required likelihood. Hence, the forward-backward procedure is used to provide solution to this problem in mathematically tractable fashion. In most HMM based speech recognition, the likelihood can be approximated by considering the best state sequence which yields the highest likelihood. This only requires the calculation of the maximum likelihood score:

\[ \hat{P}(O|\lambda) = \max_{all Q} \pi_{s_1}b_{s_1}(o_1)a_{s_1s_2}b_{s_2}(o_2) \cdots a_{s_{T-1}s_T}b_{s_T}(o_T) \]  

(3.10)

In order to do this, the Viterbi algorithm, an efficient dynamic programming method is employed by searching through the array of states for the path of greatest likelihood [56, 89].

Let a sequence of feature vectors \( O \) which is extracted from speech frames of a spoken word \( w_i \) in the decoding process. The recognition of speech is found from the optimal classification problem

\[ \hat{w}_i = \arg\max_{w_i \in W} \{ P(w_i|O) \} \]  

(3.11)

where \( \hat{w}_i \) is the hypothesised word and \( W \) is the set of all possible words that can be hypothesised by the speech recognition system. The expression above is not directly computable, and hence the Bayes rule is used to rewrite the Equation 3.11 as

\[ \hat{w}_i = \arg\max_{w_i \in W} \frac{P(O|w_i)P(w_i)}{P(O)} \]  

(3.12)
3.2. Fundamentals of Speech Recognition

where the term $P(O|w_i)$ is the acoustic likelihood and $P(w_i)$ is the language score of a particular word $w_i$. The latter term is calculated from the language model which gives the a priori probability of the occurrence of a particular word. Since the maximisation in Equation 3.12 is calculated with respect to the word $w_i$ for a given sequence of feature vectors $O$, the denominator $P(O)$ can be ignored in the maximisation. Given a HMM model $\{\lambda_i\}$ for a single word $\{w_i\}$ and assuming that $P(O|w_i) = P(O|\lambda_i)$, the maximisation in Equation 3.12 can be achieved by maximising Equation 3.10.

3.2.3 Parameter Estimation

In the training process, the HMM model parameters, $\lambda = (A, B, \pi)$ are estimated from a set of training observation sequences $O_{tr}$. The solution is to maximise the likelihood of the training observation vectors,

$$p_{tr} = \arg \max_\lambda p(O_{tr}|\lambda) \quad (3.13)$$

The maximisation problem defined in Equation 3.13 can not be analytically solved for the model parameter set that globally maximises the likelihood of the observation in a close form. Thus, the Baum-Welch algorithm is commonly employed in the HMM training of speech to determine the $\lambda$ that locally maximises $p_{tr}$. Since the algorithm requires the initial guess of $\lambda$, the initial starting point must be provided which is generally done by Viterbi training. For the full details of Baum-Welch and Viterbi training algorithms, the reader is referred to [56, 89].

3.2.4 Model Adaptation

Major performance improvement in speech recognition accuracy can be obtained when the mismatch between the environment where the model is trained and the environment where the speech recognition system operates on is minimised. This mismatch can be due to channel or additive noise effects. To achieve this, a model adaptation can be performed on a small quantity of data from the operating environment. One such technique is the Maximum Likelihood Linear Regression (MLLR) [39] which has been shown to work very well in microphone array environments [26]. The MLLR estimates the linear transformation to the mean
and the variance of state distribution of the HMM using the adaptation data to maximise the likelihood of data. For adaptation experiments in this thesis, only the mean of Gaussian is adapted by the transformation matrix:

\[ \hat{\mu} = A\mu + b \]  

(3.14)

where \( \hat{\mu} \) is the adapted Gaussian mean vector of \( \mu \) and \( A \) and \( b \) are regression factors that performs the transformation. MLLR adaptation can be performed in either a supervised or an unsupervised manner. The supervised adaptation requires adaptation data to be transcribed while the unsupervised adaptation is carried out on the data to be recognised. For further details of the MLLR technique, including the variance adaptation, the reader should refer to [39, 89].

### 3.2.5 Measuring Speech Recognition Performance

The performance of speech recogniser is evaluated by comparing the transcriptions of generated spoken word hypothesis with known reference transcriptions. The two sets of transcriptions are aligned by using an optimal string match without considering boundary timing information in the transcription. Three main types of errors occurred once the optimal alignment was found:

1. **Deletion**: This error occurred when a word from the generated transcription does not appear in the reference transcription.

2. **Substitution**: This error occurred when a word in the reference transcription is substituted by a different word in the generated transcription.

3. **Insertion**: This error occurred when a new word is inserted in the generated transcription but does not exist in the reference transcription.

Once these errors are identified after the alignment, the speech recognition performance is expressed in terms of Word Error Rate (WER) of the form [89]

\[ \%WER = \frac{D + S + I}{N} \times 100\% \]  

(3.15)

where \( N \) represents the total number of words in the experiments, \( D \) the number of deletions, \( S \) the number of substitutions, and \( I \) the number of insertions.
3.3 Microphone Array for Robust Speech Recognition

The use of multi microphones to capture a desired signal by rejecting noise and interference by means of spatial filtering or beamforming has proven to be an effective way of speech enhancement for speech recognition applications in distant-talking environments [20, 35]. Microphone arrays for robust speech recognition in general encompass all techniques for minimising the mismatch between the training data and the input speech signal through a series of enhancement procedures with the means of an array of microphones. Conventional microphone-array-based speech recognition as depicted in Figure 3.4 consists of two major parts, which are a microphone array processing block where the input speech signals are captured and enhanced, and a speech recognition system where the enhanced signal is turned into text.

The speech recogniser is usually trained on the “clean” speech collected in a quiet environment without taking into account variables that occur in real-world environments. Therefore, its performance can be expected to deteriorate when it tries to recognise noisy input signals collected in such an environment. In order to obtain the highest speech recognition performance possible, it is necessary for the system to be robust to interfering noise. Many approaches have been suggested to combat this problem. A survey by [27] identifies three main approaches relating to noisy speech recognition techniques.

The first is an attempt to recover parameter vectors of the clean speech from the corrupted waveform through a series of signal processing techniques. For
microphone array processing, this usually includes robust beamforming and post filtering [8, 46, 90]. This has been successfully applied in the past and significant performance gain is reported. While such multi-channel speech enhancement methods have been shown to improve recognition accuracy, most algorithms are optimised for noise reduction rather than for speech recognition. For some applications such as meetings, a significant reverberation could cause poor speech recognition performance. In this situation, standard beamforming may not able to negate the effects of reverberations alone and dereverberation needs to be performed. One possible extension of a dereverberation approach to multi-channel cases is proposed in [52]. While showing promising benefits, it is however inconvenient since it requires that the room transfer functions from the source to each microphone in the array to be known \textit{a priori}.

The second approach is the use of noise resistant features. Most microphone-array-based speech recognition uses MFCC or Perceptual Linear Predictive (PLP) as the accepted standard parameterisation technique. In addition to noise, the convolutive nature of channel effects can be reduced by simple manipulation in the log-spectral domain. The simplest method to do this is through \textit{cepstral mean normalisation} in which the mean of each cepstral values from an input utterance is subtracted from each cepstral value in each input frame.

The third class of methods attempts to modify the speech models to mimic the noisy environments that the recognition system operates on. The most common method employed to achieve this is by using model adaptation such as MLLR as explained in Section 3.2.4. Techniques such as Parallel Model Combination (PMC) do not require adaptation data but need statistics of the background noise. The PMC provides a framework to create single noisy HMM models from the combination of clean HMMs and noise HMMs in the log-spectral domain [25]. While the PMC has been shown to offer significant performance improvement in noisy environment, it may be not so robust in the environment where the noise characteristics are unknown or continually changing.

Other techniques which may not fall in those categories, still try to achieve some robustness to noise and at the same time attempt to integrate multi-channel processing and speech recognition systems. In multi-channel sub-band speech
recognition for example, the microphone array has been used to provide constant frequency beamforming for particular band. The sub-band speech recognition performs recognition in each band independently and the overall recognition output is obtained by combining weighted sub-band results to reach global decision. The SNR dependent weight is determined by taking advantage of multi-channel processing to estimate input signal to noise ratio [48]. Also in [50], enhanced output of the microphone array has been used to determine realibility mask in missing data speech recognition. They showed that the proposed technique gave an improved recognition result when compared to both standard beamforming output and a single channel missing data technique.

A recent unified approach to microphone-array-based speech recognition is the Likelihood Maximising Beamformer (LIMABEAM) [69, 70]. In this approach, the microphone array processing and the speech recognition system are no longer two cascaded processing blocks but instead become one integrated system. The underlying assumption behind LIMABEAM is that the recognition accuracy will be improved by determining the optimum beamforming filter which will then generate a sequence of features that maximise the likelihood of correct transcriptions. While such integration has been shown to improve speech recognition accuracy over a standard approach, it is not practical as it needs calibration to the operating environment by asking the user to speak enrollment utterances with known transcriptions. Furthermore, the algorithm is significantly more complex to implement and computationally expensive as it requires non-linear optimisation to compute array parameters.

3.4 Research Directions

While a significant amount of research has been done in the field of robust speech recognition using microphone arrays, the enhancement based approaches in which the array processing acts as a front-end pre-processing to generate a cleaner signal for the input of speech recognition system are still a popular choice in many recognition tasks [68]. Note also that many microphone arrays used in the enhancement module have a fixed, known geometry. As this thesis is focused more
on the problem of beamforming from ad-hoc arrays for speech recognition in meetings, the conventional method is still adopted and the integration between array processing and speech recognition system is considered as a separate research problem.

Meeting is a basic human communication strategy to share feelings and ideas. This occurs when at least two people meet and talk to each other. In formal meetings such as those occurring in offices or conferences, there is a need to record and analyse these as a part of the formal business strategies employed in order to increase efficiency in the workplace and enhance the productivity of these meetings. For this purpose, the upcoming meetings could be better prepared, part of meetings missed could be retrieved for review, and the participant behaviours could also be analysed [1].

Automatic speech recognition for transcribing speech in meetings poses a challenging task. As a predominant mode of interactions, speech recordings obtained during meetings are influenced by a number of factors. Such factors originate from the speaker himself as well as from environments. The speech uttered by the speaker during meetings is spontaneous and has unconstrained vocabularies. Depending on the location of microphones, there are variations in the recorded speech quality as the distance from the speaker to the each microphone is different, in which reverberations may also play apart.

To acquire a high quality signal for speech recognition, it is crucial to be able to locate and track talking people in meeting situations as meetings are often recorded with multiple distant microphones. The typical array geometries to achieve this for the current applications are linear or circular arrays in which omnidirectional microphones are used. Even though multiple arrays could be deployed in various locations to capture more complete acoustic conditions, they are usually fixed to a certain location independent of the speaker position. Another possible means to acquire data is to have huge microphone arrays mounted on the wall that essentially surround all possible sound sources. While using such immense numbers of microphones yields better speaker localisation accuracy and SNR enhancement, it is computationally complex and costly to implement [73, 74].

Motivated by a desire to progress from traditional microphone arrays towards
less constrained microphone networks, there is a potential for microphones to be deployed in an ad-hoc manner with random array geometries. Such situations can occur in the scenario where each meeting participant brings their own microphone-equipped laptop or Personal Digital Assistant (PDA) [57]. In this ad-hoc situation, localisation and tracking algorithms must be designed to work robustly with unknown microphone and speaker locations. While these are not truly ad-hoc arrays, conditions approaching this have in effect been imposed in recent NIST ASR evaluations on distant microphone recordings of meetings [21]. For example, the Computers In the Human Interaction Loop (CHIL) 2007 evaluation data provided for the Rich Transcription 2007 Meeting Recognition Evaluation (RT07) consist of 25 interactive seminars recorded at five different recording sites in Europe and the United States in multi-sensory smart rooms [10].

The AMI consortium and ICSI-Berkeley, the two major participants in the NIST evaluation, have produced their own complete system for recognising such data. While providing considerable improvement for far-field microphone processing towards speech recognition performance in close-talking microphone systems, their front-end array processing is still a conventional delay-sum algorithm [30, 79]. In order to locate and track the speaker, the steering vector for beamforming is directly estimated from the received signal delays in their both front-end system. The ICSI front-end system however has extra processing which use Viterbi algorithm to post-process delays so that only delays which correspond to the desired speaker are used for the formulation of the steering vector in the beamforming filter [4].

Based on these reviews, issues that have not received significant attention in the research literature have been identified and this thesis seeks to address them. The first limitation is the fact that the current ad-hoc array beamforming technique use the simplest beamforming algorithm, delay-sum beamforming, in spite of other existing techniques offering optimal noise reduction.

The second limitation of the existing research is the problem of speaker overlap which often occurred in meetings.

The third major limitation is whether it is best to use all microphones in such situations, or to select an optimal subset of these (e.g. the ones which are closer
to the speaker). It was reported in the AMI system that for sparsely located microphones the delay estimates are very unreliable and in this case simply selecting the channel with the highest energy was found to yield substantially lower WER [30].

This thesis aims to contribute to the field of ad-hoc microphone array processing for speech recognition in meetings by addressing these limitations. In order to address the first limitation, the next section begins with a comparative study of robust beamforming techniques when there is error in microphone placements. Such errors are likely to occur in ad-hoc array situations since microphone positions can not be obtained accurately. Subsequent to this evaluation, a novel array calibration technique is presented in Chapter 4 that is able to reveal unknown microphone geometry with reasonable position accuracy. One of the applications of the proposed technique is the possibility of dealing with speaker overlap which addresses the second limitation. In Chapter 5, the comparative study of beamforming techniques are performed for two investigated blind beamforming approaches for unknown ad-hoc array geometries. To address the third limitation, the clustering of ad-hoc microphone arrays for robust blind beamforming is proposed in chapter 6.

3.5 The Effect of Microphone Placement Errors on Recognition Accuracy

Motivated by a desire to progress from traditional microphone arrays towards less constrained microphone networks, this experiment investigates approaches that improve the robustness of a microphone array beamformer to erroneous microphone placement. Such errors or uncertainty occur when the position of microphones are estimated using array shape calibration approaches [57, 60, 61, 88].

A first approach to this problem aims to design beamformers that are inherently insensitive to sub-optimal conditions and errors, rather than correcting the errors themselves. This is usually achieved using adaptive methods to estimate and minimise the noise and steering vector uncertainties from the input
3.5. The Effect of Microphone Placement Errors on Recognition Accuracy

signals \([14, 33]\). Theoretical statistical analysis on the signal-to-interference-plus-noise ratio (SINR) performance of robust beamformers in the presence of random steering vector errors has been thoroughly presented in the literature \([5, 85]\). While such analysis provides a valuable insight into the effectiveness and optimality of the beamformer, there is a need to confirm the practical relevance of the analysis, for instance when the beamformer is used as a front-end in a speech recognition system. In such cases, eventual speech recognition accuracy is a more pertinent measure than the theoretical SINR.

The first experiment is conducted to investigate the effect of uncertain microphone placement on beamforming accuracy only and the speaker location is assumed to be known, ignoring the effects of microphone placement uncertainty on automatic speaker localisation. Five beamforming techniques which are conventional delay-sum, superdirective, MVDR, RGSC with delay-sum as the fixed beamformer component (RGSC\(_{DS}\)), and RGSC with superdirective beamformer as the fixed beamformer (RGSC\(_{SD}\)) are compared. The aim is to reveal the inherent robustness of those beamformers to sub-optimal conditions and errors.

Speech recognition experiments were conducted on the single speaker (S1) portion of the Multichannel Overlapping Numbers Corpus (MONC) \([53]\). The MONC contains digit utterances spoken around a circular meeting room table and captured by a fixed 8-element, equally spaced, table-top circular microphone array. The noise conditions in the recording was predominantly diffuse background noise. MONC was particularly chosen because it was recorded with fixed loudspeakers of known positions. Figure 3.5 show the location of microphones and loudspeakers used in the recordings.

The clean MONC S1 training data set was used to train baseline HMM with standard MFCC parameters (including 0th cepstral coefficient) and their first and second derivatives. The baseline system achieves a word error rate of 4.37% on the test set. In the following experiments, MLLR adaptation of the models is performed for each technique using the MONC development set. The WER for each robust beamformer using ground truth microphone placements is given in table 3.1.
To simulate the placement uncertainty for the robust beamforming experiments, the true microphone locations were perturbed with a random angle for a specified radius $r$. The radius of error was increased step by step, initially with a 0.2cm increment from the actual microphone positions up to 2cm, followed by 0.5cm increments from 2cm to 5cm. Due to randomisation, the results are averaged over 15 experiments for each increment of error. The mean WER results are plotted against the placement error in Figure 3.6.

At true microphone positions as presented in Table 3.1, superdirective, MVDR,

<table>
<thead>
<tr>
<th>Techniques</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delay-sum</td>
<td>7.38%</td>
</tr>
<tr>
<td>Superdirective</td>
<td>6.25%</td>
</tr>
<tr>
<td>MVDR</td>
<td>6.27%</td>
</tr>
<tr>
<td>RGSC_DS</td>
<td>6.78%</td>
</tr>
<tr>
<td>RGSC_SD</td>
<td>6.16%</td>
</tr>
</tbody>
</table>
Figure 3.6: Word error rate versus microphone placement error for various robust beamforming approaches.

and RGSC_SD achieve the lowest WER among all techniques, as the MONC database was recorded in approximately diffuse noise conditions (i.e., a moderately reverberant room with no significant localised noise sources). RGSC_DS shows significant reduction of WER compare to the standard delay and sum beamforming techniques due to the adaptive noise cancelling structure.

In the presence of small array perturbations, for radius error from 0.2cm to 2cm, all beamforming techniques exhibit relatively stable performance. This indicates a certain degree of robustness in the presence of array mismatch. As the radius of error increases, the speech recognition accuracy for all beamforming techniques begins to decline.

### 3.6 Summary

This chapter has presented a literature review on speech recognition system, starting from its fundamental theory with an overview of speech feature extraction, and the decoding and training processes. This is then followed by different approaches to microphone-array-based speech recognition systems. As this thesis is focused on the meeting task, a review of the current front-end microphone array processing for speech recognition in meetings is presented, and from this review a number of limitations in the existing research are identified.
In the first experiment of this thesis, the effect of microphone placement errors on the accuracy of a microphone array speech recognition system has been examined for various robust beamforming approaches. For the investigated configuration, results showed that beamformers were robust within 2 cm of microphone placement error. While only limited to one geometry, these experiments showed that the beamforming techniques were robust to a certain degree of placement errors without significantly affecting the speech recognition accuracy.

The research contained in this chapter is the subject of the following fully-reviewed publication:

Chapter 4

Automatic Array Calibration

4.1 Introduction

In ad-hoc situations when microphones have not been systematically positioned, the microphone position and the likely source locations are not known and beamforming must be achieved blindly. The first approach presented in this chapter is to calibrate both microphone and speaker location prior to beamforming. Naturally, calibration approaches in turn rely on other forms of prior information, such as the availability of purpose-built devices, the presence of a known calibration signal, or an assumed model of the background noise. Experiments in the previous chapter have shown that the beamforming techniques investigated are robust to a certain degree of microphone placement errors. With this in mind, the novel array calibration technique is proposed in joint collaboration with Iain McCowan and Mike Lincoln [47] which enable us to reveal microphone positions automatically and with reasonable accuracy. Speaker localisation can then be performed from the automatically derived microphone positions to enable full automatic calibration of both microphone and speaker positions. Application of the automatic calibration technique for blind source separation is then presented as a method to deal with overlapped speech in meetings.
4.2 Array Shape Calibration

A technique to calibrate or find the microphone locations within an array is also known as array shape calibration. Array shape calibration usually requires prior knowledge of initial array geometry estimates and either source position or source Direction of Arrival (DOA) depending on the signal propagation model (e.g. near-field or far-field). The purpose of the algorithm is to estimate the true values of the array parameters, which differ from the nominal values due to manufacturing errors and ageing. The technique was originally developed for antenna, radar, and sonar applications. This was motivated by the availability of high-resolution algorithms for source localisation and tracking, which although very accurate and robust in performance, are also very susceptible to errors in the array manifold.

Array shape calibration techniques deal mostly with uncertain sensor locations assuming that each sensor experiences random but time-invariant perturbations from its nominal locations. There are two approaches, the first is the self-calibrating techniques which uses sources in unknown locations [60, 87], and the second approach is based on known calibrating sources [43, 71]. The self-calibrating techniques are particularly attractive since in the real-world situation, the position of sources is rarely known.

Rockah and Schultheiss have shown that in order to self-calibrate the array, at least three spatially disjointed sources are required, and the exact location of one sensor and direction to another sensor must be known. If the direction from a known sensor to another sensor is not known, the technique is still able to calibrate the array but it will not be possible to determine the array orientation correctly. They also showed that the technique suffers ambiguity problems when the array of sensors are collinear [60, 61]. Based on the same principle, an eigenstructure-based method for direction finding with sensor gain and phase uncertainties is proposed in [87]. The technique is based on the eigendecomposition of the sample covariance matrix of received signals which fundamentally similar to a Multiple Signal Classification (MUSIC) algorithm [66]. The actual DOA and gain/phases parameters are iteratively estimated by fixing a parameter one at the time until convergence occurs. This work is further extended in [88] using a maximum
likelihood approach to achieve array shape calibration.

Following these, many other methods of calibration are proposed to solve more complicated problems. For example, Flanagan in [22] presents calibration of large sensor position errors, Stavropoulos in [78] presents an array calibration method that is capable of handling simultaneously gain, phase, location, and mutual coupling uncertainties. Nevertheless, most of these methods are only limited to the computer simulations and evaluations of their theoretical performance rather than to practical implementations. Furthermore, most of these methods are designed for narrowband signals. An extension to non-stationary wideband signals including speech most often requires large amount of computational power.

An experimental setup for the automatic calibration of microphone position and gain for large aperture microphone array is presented in [64]. The method uses short time pulses as calibration signals in known locations with synchronous timing (e.g. capture and playback occurs simultaneously). Even though it works effectively and is robust to reverberations, the method relies on the accurate measurement of time delay and on the precise knowledge of source locations.

In 2004, Raykar et al. presented techniques of calibrating the positions of ad-hoc distributed network of heterogeneous general purpose computing platforms such as laptops, PDAs, and tablets. They also assumed that there was no synchronisation among different platforms. The solution turned out to be a nonlinear minimisation which required initial estimates to reach the global maximum. For the purpose of illustrating how the nonlinear minimisation is formulated, assumptions are made that the locations of \( N \) number of microphones and \( S \) number of speakers are not known due to random time-invariant perturbations. The initial estimation of the locations are available, but the exact knowledge of their positions are not known.

The TDOA between microphone \( i \) and microphone \( j \) due to a signal originated from source \( k \) with propagation speed \( c \) is defined as:

\[
T (\{m_i, m_j\}, s_k) = \frac{|s_k - m_i| - |s_k - m_j|}{c}
\]  \hspace{1cm} (4.1)

where \( s_k \) is the position of \( k^{th} \) source and \( m_i \) and \( m_j \) are the positions of \( i^{th} \) and \( j^{th} \) microphone respectively.
In practice, TDOAs can be obtained from cross correlation between microphone pairs. The obtained TDOAs however, are corrupted due to the ambient noise and reverberations in the room. The noise corrupted TDOAs, $\tau$ can be assumed to be normally distributed, in which:

$$\tau_{kj} = T(\{m_i, m_j\}, s_k) + \theta$$  \hspace{1cm} (4.2)

where $\theta$ is the zero mean white Gaussian noise, $\theta = \mathcal{N}(0, \sigma_{kij}^2)$.

Let $\Theta$ be a $P\times1$ column vector containing parameters to be estimated which are coordinates of microphones and speakers positions. Following similar derivation in [57], the maximum likelihood estimate of $\Theta$ is found to be a minimisation of nonlinear least squares problem:

$$\hat{\Theta}_{ML} = \arg\min_{\Theta} (J_{ML}(\Theta))$$  \hspace{1cm} (4.3)

where

$$J_{ML}(\Theta) = \sum_{k=1}^{S} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{(\tau_{kij} - \tau_{kij}^{\text{actual}})^2}{\sigma_{kij}^2}$$  \hspace{1cm} (4.4)

and the $\tau^{\text{actual}}$ is the actual TDOA calculated from the microphone and speaker positions.

The cost function defined in Equation 4.3 can have multiple global minima since the TDOA depends on pairwise distances. The value of function to be minimised will not be affected by any translation and rotation of the coordinate system. In order to make the solution invariant to rotation and translation, three arbitrary nodes are selected to lie in a plane such that the first is at $(0; 0; 0)$, the second at $(x_1; 0; 0)$, and the third at $(x_2; y_2; 0)$. For two dimensions (2D), two nodes are selected to lie in a line, the first at $(0; 0)$ and the second at $(x_1; 0)$. To eliminate the ambiguity due to reflection along the Z-axis (3D), one more node is selected to lie in the positive Z-axis or for 2D, the node is selected to lie in positive Y-axis. Also, to eliminate the reflections along X-axis and Y-axis (for 3D), the nodes that form the coordinate system are assumed to lie on the positive side of the respective axes i.e $x_1 > 0$ and $y_2 > 0$.

To obtain the solution, the total number of parameters to be estimated ($P$) must be less than or equal to the number of observations ($O$). For $N$ number of microphones, $S$ number of speakers, and $R$ number of dimension, the number
of parameters to be estimated is $RN + RS - (R(R + 1)/2)$ and the number of independent observation is $(N-1)S$. Thus, the minimum number of microphones and speakers required when both are located on the plane (2D) is 6 and 3 respectively, and when both are located in the 3D space is 10 and 4 respectively, for the maximum likelihood estimation to work.

The calibration of ad-hoc nodes which was originally proposed, also deals with the lack of synchronisation. In order to calculate this, the emission and capture start times of calibrating sources are measured and included in the formulation of maximum likelihood estimator. In addition, a special device platform in which a microphone and a loudspeaker are co-located or contained within a sensor is required. This unique configuration provides an initial guess of parameters by means of a Multidimensional Scaling (MDS) algorithm. The initial estimate of coordinate locations enables the optimisation algorithm to find the global minimum and avoid getting stuck in the local minima as frequently occurred in the optimisation problem. Further details of the calibration technique can be found in [57].

The technique proposed in [6] uses an extension to classical MDS called the basis-point classical MDS algorithm by measuring only the distances between each microphone and a small number of basis points in order to calibrate the array. These distances however, must be first obtained manually using a tape measure or a similar measuring device, making them impractical for the calibration of an ad-hoc array.

Summarising from the literature reviews above, existing methods for array shape calibration require that:

(i) The array has known nominal geometry but the exact microphone positions are unknown due to perturbation. Note that the technique proposed by Raykar et al. is able to reveal the previously unknown array geometry and do calibration afterwards, but it requires a special device platform.

(ii) Calibrating signals must be available.

In the next section, a technique of array shape calibration which does not require these conditions is presented. A technique is instead benefitting from
assumed knowledge of the noise field alone.

4.3 Microphone Array Shape Calibration in Diffuse Noise Fields

Microphone array shape calibration technique in diffuse noise fields is presented in this section. This work is a joint collaboration with Iain McCowan and Mike Lincoln [47]. The technique is able to estimate the unknown location of microphones by revealing the geometrical view of the array using only a background noise signal.

A diffuse noise model is a good approximation for a noise field in practical reverberant environments [12]. Based on the theoretical sinc function given in Equation 2.44, the coherence value at any given frequency between two microphones in a diffuse noise field can be calculated from the distance between them. As an example, Figure 4.1 shows a plot of the measured coherence function along with the theoretical sinc between microphones with a distance of 0.2m. Using the diffuse noise field model, the distance between each microphone pair, \( d_{ij} \) can be estimated by fitting Equation 2.44 to the measured noise coherence from Equation 2.42 in the least-squares sense:

\[
\varepsilon_{ij}(d) = \sum_{f=0}^{fs/2} \left| \mathcal{R}\{\Gamma_{ij}(f)\} - \text{sinc}\left(\frac{2\pi fd}{c}\right)\right|^2 
\]

\[
d_{ij} = \arg \min_d (\varepsilon_{ij}(d))
\]

Given \( N \) number of microphones, there will be \( N\text{-choose}-2 \) i.e \( \binom{N}{2} \) pair-wise distances to be estimated. In the next step, classical multidimensional scaling is used to reveal the array geometry given an inter-microphone distance matrix.

The estimation of noise cross-spectral density from the short-time noise segment may deviate from the theoretical diffuse noise field model. To improve the robustness of estimates, few noise frames can be used to combine estimates over multiple frames and remove outliers. One such method detailed in [47] uses K-Means clustering to cluster noise frames into two classes, those that well match the diffuse noise model and those that do not. The noise frames that well match
4.3. Microphone Array Shape Calibration in Diffuse Noise Fields

Figure 4.1: Measured coherence function of microphone pair with a distance of 0.2m. The theoretical sinc function is plotted for true microphone distances.

the diffuse noise model tend to have a low least squares residual error. The centroid of the cluster of distance estimates with lowest residual error is selected as the final distance estimate.

4.3.1 Multidimensional Scaling

Multidimensional scaling encompasses a collection of data analysis techniques for analysing a relation between entities by revealing the hidden structure underlying the data through a geometrical representation of their relations [15, 80]. The popularity of MDS originated in psychology and has now become a general data analysis technique in a wide variety of fields.

Each entity represents a point in a multidimensional space. The points are arranged in this space so that the distances between any two points are a monotonic function corresponding to similarities (e.g. two similar objects represent two points that are close together and two dissimilar objects represent two points that are far apart).

Many variants of MDS have been developed to deal with the types of similarities, the number of similarity matrices, and the nature of the MDS models. In particular, if the similarity data are based on Euclidian distances and there is
only one similarity matrix, classical metric MDS is able to recreate the geometrical representation of the data. The steps of the classical metric MDS algorithm to estimate microphone locations are as follows:

1. Given $N$ number of microphones in $R$-dimensional space, construct the squared-distance matrix $D$ between all nodes. Each entry is $\delta_{ij} = d_{ij}^2$, where $d_{ij}$ is the Euclidean distance between microphone $i$ and $j$.

2. Construct a double centering matrix $J = I - \frac{1}{N}11^T$ where $1$ is a vector of all ones and compute the inner product matrix $B = -\frac{1}{2}JDJ$. By definition matrix $B$ is a symetric positive definite matrix and the rank of $B$ is equal to $R$.

3. Factor $B$ using a Singular Value Decomposition (SVD) as $B = U\Lambda U^T$, where $\Lambda$ is a $N\times N$ diagonal matrix of eigenvalues. The diagonal elements are arranged as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$. The columns of $U$ are the corresponding eigenvectors.

4. Extract the first $R$ eigenvalues from $\Lambda$ and their corresponding eigenvectors $U$ to form matrix $\Lambda_R$ and $U_R$. $\Lambda_R = diag(\lambda_1, \cdots, \lambda_R)$ and corresponding eigenvectors $U_R = [v_1, \cdots, v_R]$.

5. Calculate $\hat{X} = U_R\Lambda_R^{\frac{1}{2}}$, where $\hat{X}$ is the $N\times R$ matrix of microphone location coordinates.

Table 4.1 gives flying distances between Australian major cities. In this examples, entity is the city and the flying distance is a measure of similarity. Figure 4.2 gives a geometrical representation of data which reveals the approximate location of major cities in Australia.

### 4.3.2 Experimental Results

The validation of the proposed method of array shape calibration using the real noise recordings is presented in this section. The noise auto- and cross-spectral densities are typically estimated using a recursive peridogram from the noise
4.3. Microphone Array Shape Calibration in Diffuse Noise Fields

Table 4.1: Flying distances (in km) between Australian cities.

<table>
<thead>
<tr>
<th></th>
<th>Adelaide</th>
<th>Alice Springs</th>
<th>Brisbane</th>
<th>Cairns</th>
<th>Darwin</th>
<th>Melbourne</th>
<th>Perth</th>
<th>Sydney</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adelaide</td>
<td></td>
<td>0</td>
<td>1315</td>
<td>1560</td>
<td>2105</td>
<td>975</td>
<td>2580</td>
<td>660</td>
</tr>
<tr>
<td>Alice Springs</td>
<td>1315</td>
<td>0</td>
<td>1975</td>
<td>1445</td>
<td>1955</td>
<td>1300</td>
<td>1890</td>
<td>1995</td>
</tr>
<tr>
<td>Brisbane</td>
<td>1560</td>
<td>1975</td>
<td>0</td>
<td>1410</td>
<td>930</td>
<td>2855</td>
<td>1385</td>
<td>3630</td>
</tr>
<tr>
<td>Cairns</td>
<td>2105</td>
<td>1445</td>
<td>1410</td>
<td>0</td>
<td>2055</td>
<td>1678</td>
<td>2325</td>
<td>3450</td>
</tr>
<tr>
<td>Darwin</td>
<td>975</td>
<td>1955</td>
<td>930</td>
<td>2055</td>
<td>0</td>
<td>3126</td>
<td>454</td>
<td>3088</td>
</tr>
<tr>
<td>Melbourne</td>
<td>2580</td>
<td>1300</td>
<td>2855</td>
<td>1678</td>
<td>3126</td>
<td>0</td>
<td>3147</td>
<td>2659</td>
</tr>
<tr>
<td>Perth</td>
<td>660</td>
<td>1890</td>
<td>1385</td>
<td>2325</td>
<td>454</td>
<td>3147</td>
<td>0</td>
<td>2754</td>
</tr>
<tr>
<td>Sydney</td>
<td>2140</td>
<td>1995</td>
<td>3630</td>
<td>3450</td>
<td>3088</td>
<td>2659</td>
<td>2754</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1145</td>
<td>2035</td>
<td>725</td>
<td>1950</td>
<td>260</td>
<td>3150</td>
<td>698</td>
<td>3280</td>
</tr>
</tbody>
</table>

Figure 4.2: Map of Australia based on flying distances of major cities, using classical multidimensional scaling.

signal in the microphone $i$ and $j$, $v_i$ and $v_j$ as [3]:

$$\Phi_{v_i,v_j}(f) = \alpha \Phi'_{v_i,v_j}(f) + (1 - \alpha)v_i(f)v_j(f)$$  

(4.7)

where $\Phi$ is the density estimate for the current frame, and $\Phi'$ is the estimate from the previous frame. The term $\alpha$ is a number close to unity which is given by $\alpha = \exp(-T/\tau_\gamma)$, where $T$ is the step size in seconds, and $\tau_\gamma$ is the decay time constant. The algorithm is implemented with a frame length of 64 ms, with $\tau_\gamma$ factor of 64 ms and sampling rate of 16 kHz. The curve fitting of the sinc function was done with *lsqnonlin* function from Matlab optimisation toolbox.

To quantify the accuracy of the estimated array geometry, three background noise recordings of 10 seconds duration were used.
1. G1: The noise was recorded using an equispaced (5cm) linear array of eight microphones. The noise mostly comes from computers, recording equipments, projectors, and air conditioners.

2. G2: The noise was recorded in identical condition to G1, using an eight element square microphone array with a side length of 0.2m.

3. G3: The noise was recorded using an equally spaced, circular array with a diameter of 0.2m in similar condition to G1.

The accuracy of array calibration estimates are presented in two different forms:

(i) Accuracy of inter-microphone distance estimates: The inter-microphone distance error $\delta_d$ is calculated for each pair of microphones. For $N$ microphones, the mean, standard deviation, minimum, and maximum error of $\frac{N(N-1)}{2}$ distance estimates are shown in Table 4.2.

(ii) Accuracy of microphone location estimates: The MDS algorithm only determines the relative geometry of the microphones since it only depends on pairwise distances. An optimal rotation, translation, and reflection need to be calculated to align the relative geometry to the actual microphone positions in order to calculate the accuracy of position. This type of matching is usually known of as procustes analysis which uses a goodness-of-fit criterion to minimise the sum of squared errors between the two configurations. 
*Procrustes* function in Matlab statistic toolbox can be used to do this analysis. The location error $\delta_p$ of each microphone is calculated and the overall mean, standard deviation, minimum, and maximum error of $N$ microphones are shown in Table 4.3.

The geometry of the calibrated microphone positions for linear, square, and circular microphone arrays are shown in Figure 4.3.

The performance of array shape calibration shows an overall good accuracy of microphone location estimates with an average error less than 0.7cm with the maximum error only within 2cm for square and circular array. For linear array,
Figure 4.3: Actual (circles) and its corresponding estimated microphone location (pluses) for linear array (top), square array (middle), and circular array (bottom).
Table 4.2: Accuracy of inter-microphone distance estimates of the calibrated array (in cm).

<table>
<thead>
<tr>
<th>Recording</th>
<th>$\delta_d$</th>
<th>$\sigma_d$</th>
<th>$\min(\delta_d)$</th>
<th>$\max(\delta_d)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>0.13</td>
<td>0.15</td>
<td>0.00</td>
<td>0.50</td>
</tr>
<tr>
<td>G2</td>
<td>0.56</td>
<td>0.45</td>
<td>0.00</td>
<td>1.92</td>
</tr>
<tr>
<td>G3</td>
<td>0.36</td>
<td>0.31</td>
<td>0.01</td>
<td>1.04</td>
</tr>
</tbody>
</table>

Table 4.3: Accuracy of microphone location estimates of the calibrated array (in cm).

<table>
<thead>
<tr>
<th>Recording</th>
<th>$\delta_p$</th>
<th>$\sigma_p$</th>
<th>$\min(\delta_p)$</th>
<th>$\max(\delta_p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>2.09</td>
<td>1.69</td>
<td>0.47</td>
<td>4.00</td>
</tr>
<tr>
<td>G2</td>
<td>0.68</td>
<td>0.31</td>
<td>0.34</td>
<td>1.26</td>
</tr>
<tr>
<td>G3</td>
<td>0.65</td>
<td>0.31</td>
<td>0.26</td>
<td>1.06</td>
</tr>
</tbody>
</table>

the small error in inter-microphone distance estimates show that the distances between each microphone are accurately estimated. However, the output of multidimensional scaling gives small deviation in position if the array is actually linear as $R$ was set equal to 2, resulting large errors in location estimates.

Using more noise frames may further improve the robustness of estimates as [47] found, nevertheless care must be taken that no speech frames or higher energy noise frames are taken as they may contain other localised noise sources in addition to the diffuse background.

Figure 4.4 shows the spatial response plot for delay-sum and superdirective beamformer with weights calculated using true microphone positions and estimated positions. The plot shows that the errors in the calibrated position do not change the look direction of the array, showing that the calibration approach is quite robust for these shapes of the array.

### 4.4 Speaker Localisation

Speech applications such as video conferencing requires the microphone array and camera to track the moving speaker to keep the speaker within focus of the camera. In the multi-party meeting scenario for example, the localisation of the
speaker is necessary to enhance the speech signal by steering the microphone array to a particular speaker with respect to others. To enable this to occur, speaker localisation is necessary in the development of speech enhancement methods that require the speaker position to be known. Conventional beamforming needs accurate location of both microphone and speaker positions for its optimal use since beamformer will pick up the noise or unwanted sound sources if there is inaccuracy in the steering direction. Existing source localisation algorithms can be loosely classified based on their method of estimation into three categories [18]:

(A) Steered Beamformer Approach: This approach relies on steering the beamformer’s array response to various locations of interest and searching the peak in the output power [11, 86]. To find the global maximum, the maximum likelihood location estimator is usually formulated. While this tends to be robust, the physical realisation of the maximum likelihood estimator...
involves nonlinear optimisation and therefore a good initial source of location estimates is needed to avoid getting stuck in the local maximum. Furthermore, this approach is highly dependent on the spectral content of source signal which is not practical in most applications.

(B) High Resolution Spectral Analysis: The second category adapts techniques from the field of spectral analysis to do source localisation. The high resolution spectral analysis mostly relies on the decomposition of the cross spectral density matrix of signals received at the sensors. When the exact knowledge of this matrix is unknown (as in most cases), the matrix is obtained via the ensemble averaging of the observed signals over an interval in which they are assumed to be stationary with the estimation parameters being fixed. This can be problematic for speech signals in practice. These techniques are originally designed for a narrowband signal and extending them to wideband will involve extra computational power. The techniques are also not robust to modelling errors, and thus are very sensitive to error in sensor position or changes in channel characteristics, making them impractical to be implemented in real-world environments.

(C) TDOA Based Techniques: The last category employs two independent steps to determine source location. Firstly, the TDOA is estimated between pairs of spatially separated microphones, and then by utilising the knowledge of sensor positions, the hyperbolic curves are generated and intersected in the optimal sense to arrive at the source location. A number of techniques which have closed form solutions have been developed based on this principle [9, 65, 76].

From those categories, the TDOA based approach is widely used in practice due to its simplicity and this works effectively with speech as an input signal. In this thesis, Steered Response Power with Phase Transform weighting (SRP-PHAT), a technique which combines the advantage of steered beamformer with the TDOA approach is used for speaker localisation. SRP-PHAT has been shown to produce more accurate and robust results than other classes in many research papers.
4.4. Speaker Localisation

4.4.1 SRP-PHAT Localisation

The SRP localisation computes the steered response of the beamformer by summing the cross-correlation of all pairs of microphones in the array. The cross correlation between the signal at microphone $i$, $x_i$, and $j$, $x_j$ also calculates the TDOA between these two microphones. This is given by:

$$ c^{ij}(\tau) = \int_{-\infty}^{\infty} x_i(t)x_j(t-\tau)dt \quad (4.8) $$

In the frequency domain by means of Fourier transform, the cross correlation is known as cross spectral density which is defined as

$$ C^{ij}(f) = X_i(f)X_j^*(f) \quad (4.9) $$

where $X_i(f)$ is the Fourier transform of $x_i(t)$ and $X_j^*(f)$ is the complex conjugate of the Fourier transform of $x_j(t)$.

The inverse Fourier transform gives cross correlation in terms of $X_i(f)$ and $X_j(f)$ as:

$$ c^{ij}(\tau) = \int_{-\infty}^{+\infty} X_i(f)X_j^*(f)e^{j2\pi f\tau} df \quad (4.10) $$

Due to the noise affecting the peak estimate, it is desirable to filter the signal prior to the cross correlation [38]. Introducing $G_i(f)$ and $G_j(f)$ as a frequency domain filter to filter signal $X_i(f)$ and $X_j(f)$ respectively, the Generalised Cross Correlation (GCC) can be formulated as

$$ c^{ij}_{GCC}(\tau) = \int_{-\infty}^{+\infty} \Psi_{ij}(f)X_i(f)X_j^*(f)e^{j2\pi f\tau} df \quad (4.11) $$

where $\Psi_{ij}(f) = G_i(f)G_j^*(f)$ is the weighting function.

If $\Psi_{ij}(f) = 1$, the standard cross correlation formula is obtained. Various weighting functions however can be used such as the Maximum Likelihood (ML), Smoothed Coherence Transform (SCOT), and Eckart. One weighting function which has been shown to perform well in moderately noisy and reverberant environments and which is widely used in practice is the Phase Transform (PHAT) weighting [38]. The PHAT weighting is defined as

$$ \Psi_{ij}(f) = \frac{1}{|X_i(f)X_j^*(f)|} \quad (4.12) $$
The PHAT places equal emphasis on each component of the cross spectrum phase by flattening out the magnitude spectrum. The PHAT-weighted cross correlation between microphone $i$ and $j$, $c^{ij}_{\text{GCC-PHAT}}(\tau)$ is given by:

$$c^{ij}_{\text{GCC-PHAT}}(\tau) = \int_{-\infty}^{+\infty} \frac{X_i(f)X^*_j(f)}{|X_i(f)X^*_j(f)|} e^{j2\pi f \tau} df$$  \hspace{1cm} (4.13)

The $c^{ij}_{\text{GCC-PHAT}}(\tau)$ will exhibit a global maximum at the lag value which corresponds to the relative delay which is the TDOA between microphone $i$ and $j$:

$$\text{TDOA}^{ij} = \arg \max_{\tau} \left(c^{ij}_{\text{GCC-PHAT}}(\tau)\right)$$  \hspace{1cm} (4.14)

For the purpose of source localisation, the SRP-PHAT of a delay-sum beamformer directed at location $S$ can be shown to be equivalent to the sum of the PHAT weighted GCC’s for all pairs of microphones within the array:

$$P(S) = \sum_{i=1}^{N} \sum_{j=1}^{N} c^{ij}_{\text{GCC-PHAT}}(\delta(S, i, j))$$  \hspace{1cm} (4.15)

where $c^{ij}_{\text{GCC-PHAT}}(\tau)$ has been given in Equation 4.13 and $\delta(S, i, j)$ is the time delay of arrival between microphones $i$ and $j$ for a source located at position $S$,

$$\delta(S, i, j) = \frac{|S - L_i| - |S - L_j|}{c}$$  \hspace{1cm} (4.16)

The $L_i$ and $L_j$ are the locations of microphones $i$ and $j$. In the single speaker scenario, a search is performed over a number of locations $S$ and the speaker’s location is taken as the value of $S$ which maximises $P(S)$.

### 4.5 Automatic Calibration

Using the proposed array shape calibration technique to reveal microphone positions followed by SRP-PHAT to find the speaker location from the estimated microphone locations, the full automatic array calibration can be performed given only the microphone array recordings. The steps of automatic calibration are outlined below:

1. First, the noise segments are extracted from the recordings using a voice activity detector. The typical single feature energy-based speech detector
can be employed to discriminate between the background noise segment and the speech segments.

2. Microphone locations are then estimated using the array shape calibration technique presented above using the background noise segment.

3. Utilising the knowledge of microphone positions, the SRP-PHAT is performed using the speech segments to find the speaker positions.

It is possible to further calibrate both microphone and speaker positions jointly using the maximum likelihood estimator in Equation 4.3. In order to do this, the estimated microphone and speaker positions from automatic calibration are used as initial guesses for an optimisation function. The solutions can only be obtained if the number of speakers and microphones that are required are satisfied.

### 4.5.1 Experimental Results

To quantify the effect of the calibration error on the source localisation performance, the estimated source positions were compared with the ground truth source locations obtained from SRP-PHAT using calibrated and true array geometry. The evaluations were carried using test recording from the MONC database. The MONC database contains digit utterances spoken by a fixed known speaker position around a table-top circular microphone array. The array and speaker configuration is shown in Figure 3.5.

The accuracies of calibrated microphone locations from MONC recording is given in Table 4.4 and Table 4.5. The calibration error of the circular array in the MONC recording is slightly larger than the one used in the previous section to validate the array shape calibration technique. Due to the large error in the calibrated array geometry in MONC, an additional source localisation experiment was also carried out by using the geometry of the calibrated array used to validate the array shape algorithm in Figure 4.3(c).

Given the centre microphone as the origin, the speaker S1 is located in the x-axis $x = 0.495m$, y-axis $y = -0.495m$, and z-axis $z = 0.35m$. In the spherical
Table 4.4: Accuracy of inter-microphone distance estimates of the calibrated circular array from MONC recording (in cm).

<table>
<thead>
<tr>
<th>$\delta_d$</th>
<th>$\sigma_d$</th>
<th>$\min(\delta_d)$</th>
<th>$\max(\delta_d)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.10</td>
<td>1.93</td>
<td>0.09</td>
<td>7.45</td>
</tr>
</tbody>
</table>

Table 4.5: Accuracy of microphone location estimates of the calibrated circular array from MONC recording (in cm).

<table>
<thead>
<tr>
<th>$\delta_p$</th>
<th>$\sigma_p$</th>
<th>$\min(\delta_p)$</th>
<th>$\max(\delta_p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.53</td>
<td>1.17</td>
<td>0.46</td>
<td>4.15</td>
</tr>
</tbody>
</table>

coordinates, S1 is located in the azimuth $\phi = -45^\circ$, elevation $\theta = 26.56^\circ$, and radius $r = 0.7m$. Table 4.6 shows the localisation errors of the speaker S1 in terms of azimuth (degrees), elevation (degrees), and radius (meters) for both ground truth microphone locations and those estimated using the procedure described in Section 4.3. The errors are averaged over 2061 estimates and correspond to the total number of files in the MONC testing set.

The ground truth microphone positions and the calibrated 2 array positions produce a good accuracy of speaker position estimates, with the exception of the large inaccuracy in the elevation angle. The large error in the calibrated MONC array (calibrated 1) gives inaccurate speaker positions far from the true location. This result suggests that accurate calibrated microphone positions will yield small errors in speaker position estimates.

The speech recognition experiments for the automatic calibration technique

Table 4.6: Accuracy of speaker localisation results in terms of azimuth (degrees), elevation (degrees), and radius (metres) error using the ground truth array geometry (Ground truth), calibrated array from MONC (Calibrated 1), and calibrated array from Figure 4.3(c). (Calibrated 2).

<table>
<thead>
<tr>
<th>array</th>
<th>azimuth mean</th>
<th>azimuth stdev</th>
<th>elevation mean</th>
<th>elevation stdev</th>
<th>radius mean</th>
<th>radius stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground truth</td>
<td>1.89</td>
<td>1.38</td>
<td>1.83</td>
<td>5.08</td>
<td>0.10</td>
<td>0.07</td>
</tr>
<tr>
<td>Calibrated 1</td>
<td>9.94</td>
<td>2.50</td>
<td>14.24</td>
<td>2.47</td>
<td>0.16</td>
<td>0.09</td>
</tr>
<tr>
<td>Calibrated 2</td>
<td>0.86</td>
<td>1.29</td>
<td>6.08</td>
<td>3.66</td>
<td>0.13</td>
<td>0.07</td>
</tr>
</tbody>
</table>
will be studied in the next chapter. This will then be compared with the direct estimation technique for blind beamforming from ad-hoc arrays. The meeting task of the medium-sized vocabulary will be used as a test set for the comparison between the two algorithms.

4.6 Beamforming Approach to Blind Speech Separation

Recently there has been a large effort to produce automatic speech recognition systems which can automatically transcribe multi-party conversations such as those which occur during meetings [30, 37]. Overlapping speech segments, where more than one individual is talking simultaneously, pose a serious problem for such systems, increasing the absolute word error rate between 15-30% when using close talking microphones for a large vocabulary task [54]. However most systems are optimised for a single dominant speaker, with performance degrading significantly when competing speech is present. One of the reasons for this is that, while a number of different algorithms exist for separating the speech of competing talkers, to date there has been no systematic comparison of the approaches in terms of their effects on recognition performance. Algorithms are typically evaluated in terms of signal to noise ratio on small datasets containing signals in the presence of noise or other interfering signals, and these datasets are often specific to the group conducting the research, making comparison of algorithms difficult.

The PASCAL SSC2 [42] attempts to overcome this lack of comparable results by providing a unified framework in which speech separation algorithms may be compared in terms of speech recognition word error rate on a standardised task, with a baseline recogniser. The task consists of separating the overlapping speech of two talkers simultaneously reading sentences from the Wall Street Journal (WSJ) database in an instrumented meeting room. The speakers are recorded using headset and lapel microphones and two 8-element circular microphone arrays placed on a table in the centre of the meeting room. A development set consisting of 178 sentences and an evaluation set of 143 sentences are provided.
along with a baseline WSJ recogniser capable of recognising the separated speech.

There are two common approaches to the separation of overlapping speech: Blind Source Separation (BSS) and microphone array processing. Blind source separation attempts to exploit the assumption of statistical independence between the overlapped signals in order to separate them, while microphone array processing provides an enhanced version of the input speech based on the location of the speakers. Microphone array processing techniques generally require \textit{a priori} knowledge of the microphone and speaker locations, while BSS algorithms typically do not. In this section, an automatic calibration technique for microphone array beamforming is proposed as an alternative to blind source separation. The technique automatically generates the required speaker and microphone location information which is therefore directly comparable to BSS approaches.

4.6.1 System Overview

The system consists of four sub-components as shown in Figure 4.5. Initially in step 1, the microphone locations are estimated using the proposed array shape calibration technique in a diffuse noise field. An SRP-PHAT based localisation is then used to estimate the location of the speakers in the room in step 2. The microphone and speaker location information is used in a standard superdirective beamformer to separate the speech of the two speakers, and finally the crosstalk between the speakers is canceled by means of a masking post-filter applied to the beamformer output. The separated signals are then recognised by the ASR system for evaluation.
Step 1: Estimating Microphone Positions

The array calibration follows the procedure described in Section 4.3. Initial experiments on the SSC2 development data showed that the microphone calibration system was unable to precisely estimate the microphone positions when calibrating both arrays simultaneously, however microphones from the same array were successfully clustered near each other. This is due to the larger distances separating the two arrays - approximately 1.5 meters, compared to intra-array spacings of several centimetres. While the diffuse coherence model still holds for larger distances, it is difficult to obtain precise curve fitting as the width of the sinc main lobe approaches zero, leading to inaccurate estimates. To overcome this, a two pass procedure was used:

1. Find the initial position of all of the microphones using the calibration procedure.

2. Cluster the microphones into $K$ sub-arrays using the K-Means clustering. $K$ is initialised to 1 and incremented until no microphone within a cluster is more than 5cm from any other microphone within that cluster.

3. Re-calibrate the positions of the microphones within each sub-array.

4. Since the MDS algorithm estimates locations subject to arbitrary rotation and scaling, each sub-array is aligned to the original estimates from step 1 to ensure a common basis for all microphones.

Step 2: Estimating Speaker Positions

Rather than using both arrays, each array (sub-array) is treated as a stand-alone array for speaker localisation in order to reduce the amount of computational power due to the large search space when both arrays are used simultaneously. Using the sub-array’s calibrated microphone positions, two speaker positions are estimated using SRP-PHAT. The four estimated speaker locations (2 estimates for each sub-array) are then post-processed to obtain the final two speaker positions.
Step 3: Superdirective Beamforming

Superdirective beamformers is used to provide spatial discrimination between the two speakers. A set of superdirective channel filters are calculated for each speaker, minimising the diffuse noise power received by the sensors under the constraint of unity gain in the desired direction.

Step 4: Cross-talk Noise Cancelling Post-filter

In order to further eliminate the signal from competing speakers, a frequency domain masking post-filter is applied to each beamformer output. The post-filter is motivated by the assumption that, at each time frame, the energy spectrum of the sum of the two signals can be approximated by the maxima of the individual spectra for each signal in each frequency bin. This is due to the fact that two independent signals rarely contain high energy in the same frequency bin at the same time, a property which has been exploited to develop a series of techniques for speech separation using a single microphone [63].

If \( b_s(f) \) is the frequency domain output of beamformer \( s \), \( z_s(f) \) is the post-filtered output and \( h_s(f) \) is the frequency response of the post-filter, then for \( S \) beamformers directed at competing sources [44],

\[
z_s(f) = b_s(f) h_s(f)
\]

where

\[
h_s(f) = \begin{cases} 
1 & \text{if } s = \arg \max_{s'} |b_{s'}|, s' = 1...S \\
0 & \text{otherwise}
\end{cases}
\]

In the case of the SSC2 system with \( S = 2 \) the output of the beamformer is set to zero in a given frequency bin if the output of the second beamformer is greater in that bin.

4.6.2 Experimental Results

The experimental results for each step in the system diagram is presented below:
Step 1: Array Calibration Performance

In order to evaluate the performance of the array calibration technique on the SSC2 data, the inter-microphone distances are compared to the known ground truth microphone locations. Table 4.7 shows the mean, standard deviation, minimum and maximum errors of the inter-microphone distance estimates for each sub-array.

<table>
<thead>
<tr>
<th>sub-array</th>
<th>$\delta_d$</th>
<th>$\sigma_d$</th>
<th>$\min(\delta_d)$</th>
<th>$\max(\delta_d)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.80</td>
<td>0.65</td>
<td>0.02</td>
<td>2.33</td>
</tr>
<tr>
<td>2</td>
<td>0.46</td>
<td>0.31</td>
<td>0.02</td>
<td>1.15</td>
</tr>
</tbody>
</table>

The MDS algorithm returns the microphone locations subject to arbitrary rotation and translation. So that the location estimates may be evaluated, for each sub-array an incremental search is performed to align the estimated locations to those of the ground truth array geometry. Figure 4.6 shows the actual and estimated microphone positions after the search, and Table 4.8 gives the mean, standard deviation, minimum and maximum position errors in the microphone location estimates.

<table>
<thead>
<tr>
<th>sub-array</th>
<th>$\delta_d$</th>
<th>$\sigma_d$</th>
<th>$\min(\delta_d)$</th>
<th>$\max(\delta_d)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.69</td>
<td>0.75</td>
<td>0.01</td>
<td>2.18</td>
</tr>
<tr>
<td>2</td>
<td>0.42</td>
<td>0.28</td>
<td>0.00</td>
<td>0.82</td>
</tr>
</tbody>
</table>

The inter-microphone distance and position errors for sub-array 1 are about approximately twice that of sub-array 2, however, the microphone positions estimates for both sub-arrays show good accuracy, with errors ranging from 0cm to 2.18cm.
Step 2: Source Localisation Performance

To assess the performance of the source localisation system, the estimated source positions were compared with the ground truth source locations obtained from the multi-channel Wall Street Journal audio visual (MC-WSJ-AV) data collection documentation [41]. The documentation gives the location of the speaker in terms of their seating position around the meeting room table and from this, and from the measurements of the table size, the approximate location of the speaker can be inferred. Tables 4.9 and 4.10 show the localisation errors in terms of azimuth (degrees), elevation (degrees), and radius (meters) for both ground truth microphone locations and for those estimated using the calibration procedure. The results show that, for sub-array 1, the use of the estimated microphone positions has little effect on the ability to locate both speakers, however for sub-array 2 there is a significant decrease in the number of files in which locations for both speakers are found. This may most likely be attributed to the fact that sub-array 2 is typically further from the speakers than sub-array 1.

To improve the speaker localisation performance, location estimates from the two arrays are calculated independently and then combined as follows - since sub-array 1 provides better localisation performance than sub-array 2, the location corresponding to the most dominant peak in sub-array 1’s GCC-PHAT is taken as the location of speaker 1. For the second speaker, if estimate 2 from sub-array 1 is more than 22.5 degrees away from estimate 1 in the azimuth, then it is retained as the second location estimate, otherwise the location estimate from sub-array
### Table 4.9: % Accuracy of speaker localisation results in the development set using actual microphone positions, in terms of azimuth (degrees), elevation (degrees), and radius (metres) error.

<table>
<thead>
<tr>
<th>sub-array</th>
<th>found both</th>
<th>speaker</th>
<th>azimuth mean</th>
<th>azimuth stdev</th>
<th>elevation mean</th>
<th>elevation stdev</th>
<th>radius mean</th>
<th>radius stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(total of 178)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>134</td>
<td>1</td>
<td>4.36</td>
<td>6.84</td>
<td>7.84</td>
<td>4.82</td>
<td>0.41</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>4.59</td>
<td>5.03</td>
<td>8.21</td>
<td>5.11</td>
<td>0.48</td>
<td>0.36</td>
</tr>
<tr>
<td>2</td>
<td>143</td>
<td>1</td>
<td>8.66</td>
<td>6.57</td>
<td>7.42</td>
<td>5.67</td>
<td>0.35</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>10.33</td>
<td>24.06</td>
<td>8.24</td>
<td>8.98</td>
<td>0.37</td>
<td>0.28</td>
</tr>
</tbody>
</table>

### Table 4.10: % Accuracy of speaker localisation results in the development set using estimated microphone positions, in terms of azimuth (degrees), elevation (degrees), and radius (metres) error.

<table>
<thead>
<tr>
<th>sub-array</th>
<th>found both</th>
<th>speaker</th>
<th>azimuth mean</th>
<th>azimuth stdev</th>
<th>elevation mean</th>
<th>elevation stdev</th>
<th>radius mean</th>
<th>radius stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(total of 178)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>133</td>
<td>1</td>
<td>3.53</td>
<td>2.60</td>
<td>4.83</td>
<td>3.95</td>
<td>0.41</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>5.43</td>
<td>12.04</td>
<td>9.30</td>
<td>9.13</td>
<td>0.77</td>
<td>0.43</td>
</tr>
<tr>
<td>2</td>
<td>115</td>
<td>1</td>
<td>8.16</td>
<td>7.53</td>
<td>8.15</td>
<td>6.37</td>
<td>0.32</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>9.80</td>
<td>7.24</td>
<td>6.72</td>
<td>7.11</td>
<td>0.36</td>
<td>0.32</td>
</tr>
</tbody>
</table>
Chapter 4. Automatic Array Calibration

Table 4.11: % Accuracy of the merged speaker localisation results in the development set using actual and estimated microphone positions, in terms of azimuth (degrees), elevation (degrees), and radius (metres) error.

<table>
<thead>
<tr>
<th>geometry</th>
<th>found both (total of 178)</th>
<th>speaker</th>
<th>azimuth mean</th>
<th>azimuth stdev</th>
<th>elevation mean</th>
<th>elevation stdev</th>
<th>radius mean</th>
<th>radius stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>153</td>
<td>1</td>
<td>4.36</td>
<td>6.84</td>
<td>7.84</td>
<td>4.82</td>
<td>0.41</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>4.78</td>
<td>5.09</td>
<td>7.82</td>
<td>5.05</td>
<td>0.43</td>
<td>0.35</td>
</tr>
<tr>
<td>Calibrated</td>
<td>148</td>
<td>1</td>
<td>3.53</td>
<td>2.60</td>
<td>4.83</td>
<td>3.95</td>
<td>0.41</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>6.73</td>
<td>12.06</td>
<td>9.63</td>
<td>8.92</td>
<td>0.71</td>
<td>0.48</td>
</tr>
</tbody>
</table>

2 with the greatest angular azimuth separation from estimate 1 of sub-array 1 is selected. This technique attempts to avoid the problem of localising to the same source twice. Localisation error for the combined estimates are shown in Table 4.11 for both estimates and also for ground truth microphone locations. The results show that combining the estimates from the two arrays leads to an improvement in the speaker detection rates in addition to an improvement in the azimuth estimates for the located speakers.

Step 3 and 4: Superdirective Beamforming and Post-filtering

The spectrogram of the enhanced speech outputs as a result of speech separation using microphone arrays are plotted below along with the corresponding binary mask used in the cross-talk noise cancelling post-filter. The original overlapped speech file is a female speaker saying "If he had he probably would have informed the white house by now so that the administration could find a successor", while in the same time a male speaker saying "The matter of a public offering was discussed at the firm’s annual partners gathering here saturday". Figure 4.7 and 4.8 show the spectrogram of the separated speech output for the female and male speaker respectively. For comparison, the corresponding spectrogram of headset microphone signal for each utterance is included.

Figure 4.7: The spectrogram of the original overlapped speech, enhanced superdirective and post-filter output, and the headset microphone recording of the female speech. The corresponding binary mask for the post-filter is also plotted.
Figure 4.8: The spectrogram of the original overlapped speech, enhanced superdirective and post-filter output, and the headset microphone recording of the male speech. The corresponding binary mask for the post-filter is also plotted.
4.6.3 Speech Recognition Experiments

Speech recognition experiments were conducted on the final post-filtered output of the separation system using the recogniser supplied by the SSC2 challenge organisers [42]. The system is an HTK based HMM recogniser [89] with acoustic models trained on the WSJCAM0 database [59], a dictionary derived from that used in the AMI large vocabulary continuous speech recognition system [30] and the ARPA bigram and trigram language models developed for the WSJ0 corpus. Because the acoustic models are trained on data from close talking microphones, MLLR is used to adapt the models to the output of the separation system. Due to the small amounts of data available for adaptation, a leave one out cross-validation approach is used for adaptation.

The supplied recognition system also includes scoring scripts which automatically assign the separated speech to the correct speaker. To do this, the scoring script compares both sentences against transcripts for each of the read utterances, and selects the one which gives the highest score. Tables 4.12 and 4.13 show the recogniser word error rates for the development and evaluation sets for a number of different conditions. Headset and lapel microphones were worn by the speakers during the database recording and the results from these microphones are given, in addition to a single distant microphone condition in which a single microphone from sub-array 1 is used. ‘GT-Beamformer’ and ‘GT-Post-filter’ indicate results from using ground truth microphone and speaker locations for the beamformer both without and with the post-filter. Finally ‘Beamformer’ and ‘Post-filter’ results are given for the full system without and with post-filter, but using automatically derived microphone locations. Results are presented for the combined score for both speakers (the ‘Both’ column) and for scoring only the best scoring speaker for each utterance. This shows that in many cases one speaker is significantly easier to recognise than the other - possibly because they are either louder, clearer speakers, or they read longer sentences.

In the development set, using automatically derived microphone and speaker locations, results in slightly lower WER compared with the use of ground truth data. While this result is encouraging it should be noted that the ground truth speaker locations are only approximate (as explained in Section 3.6.2) and that
Chapter 4. Automatic Array Calibration

Table 4.12: % word error rate on development data set of the proposed system.

<table>
<thead>
<tr>
<th></th>
<th>Both</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>Headset Microphone</td>
<td>11.7</td>
<td>5.2</td>
</tr>
<tr>
<td>Lapel Microphone</td>
<td>32.2</td>
<td>20.5</td>
</tr>
<tr>
<td>Single Distant Microphone</td>
<td>98.8</td>
<td>92.5</td>
</tr>
<tr>
<td>GT Beamformer</td>
<td>68.8</td>
<td>47.1</td>
</tr>
<tr>
<td>GT Postfilter</td>
<td>57.0</td>
<td>36.2</td>
</tr>
<tr>
<td>Beamformer</td>
<td>66.3</td>
<td>43.4</td>
</tr>
<tr>
<td>Postfilter</td>
<td>54.8</td>
<td>35.1</td>
</tr>
</tbody>
</table>

Table 4.13: % word error rate on evaluation data set of the proposed system.

<table>
<thead>
<tr>
<th></th>
<th>Both</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>Headset Microphone</td>
<td>16.4</td>
<td>8.4</td>
</tr>
<tr>
<td>Lapel Microphone</td>
<td>46.3</td>
<td>26.5</td>
</tr>
<tr>
<td>Single Distant Microphone</td>
<td>103.3</td>
<td>91.0</td>
</tr>
<tr>
<td>Beamformer</td>
<td>73.9</td>
<td>50.7</td>
</tr>
<tr>
<td>Postfilter</td>
<td>58.0</td>
<td>38.3</td>
</tr>
</tbody>
</table>
the results would be expected to improve if the ground-truth speaker locations were known more accurately. The beamforming system gives much higher WER than close talking microphones, however it is closer in performance to the lapel microphones, with which microphone array systems have previously been competitive [53]. A similar pattern of results is seen in the evaluation set, which word error rates being even higher than for the development data, indicating that this is a difficult task.

An approach to speech separation based on microphone array beamforming techniques which avoids the usual problems of beamforming (that of requiring detailed knowledge of the recording environment) by automatically estimating microphone and speaker locations have been presented. The technique achieves closer performance to the lapel microphones and shows large improvements in recognition accuracy compared with the single distant microphone case. The data supplied with for the PASCAL SSC2, in which the algorithm was evaluated, is a ‘worst case scenario’ in terms of speaker overlap in that every recording contains overlapped speech. This may explain the difference between our approach and the recognition results using lapel microphones with which, in other task, beamforming approaches are usually competitive.

There are several areas in which the system could be improved. The array calibration algorithm relies on the microphones being in a diffuse noise field. While this is a good model for many practical situations it may not always be appropriate to assume the diffuse noise case, particularly when the microphones are widely spaced. In Chapter 6, the clustering technique will be investigated further in order to ensure that the calibration is not affected by widely spaced microphones.

4.7 Summary

This chapter begins with a literature review on array calibration techniques. In particular, the array shape calibration or calibration of microphone positions was detailed. One particular technique described in detail was that ad-hoc calibration of sensor networks by Raykar et al. The novel array shape calibration was
presented based on the shortcomings of the state-of-the-art techniques in the literatures. The technique was validated using the background meeting room noise and was shown to be robust with an approximately 1cm position error for small arrays. Once the estimated microphone positions were found, the SRP-PHAT localisation can be performed to obtain the speaker location. The two-step cascaded approach of finding microphone and speaker positions can be used for automatic array calibration using both noise and speech segments of microphone array recordings without \textit{a priori} information about the array geometry or speaker location. In this case, beamforming can be done from the the steering vector derived from the estimated positions.

One application of automatic array calibration proposed in this chapter is the separating of the overlapped speech using microphone arrays. In the experiments presented on the overlapped speech database for the PASCAL SSC2 challenge, the system achieves closer performance to the lapel microphones and shows large improvements in recognition accuracy compared with the single distant microphone case.

The research contained in this chapter is the subject of the following fully-reviewed publications:


Chapter 5

Blind Beamforming

5.1 Introduction

In ad-hoc situations, when microphones have not been systematically positioned, the microphone positions and the likely source locations are not known and beamforming must be achieved blindly. This scenario occurs naturally when meeting interactions are captured from a number of sites, each with their own microphone configurations. To provide transcriptions for meetings, there are two possible approaches for the estimation of the steering vector for beamforming. The first is an automatic calibration to determine microphone’s placement and speaker location, and the second is the direct estimation of the steering vector without regard to the microphone and source locations.

Such a direct estimation approach has been used for the Multiple Distant Microphone (MDM) condition in the AMI system [30] and the ICSI system[21, 37], among others in the NIST meeting data evaluations. While this approach has been shown to be effective, it is based on a simple delay-sum beamforming algorithm which is generally a sub-optimal solution in a practical noise environments. Other beamformers such as superdirective and MVDR have shown superior performance over delay-sum in the meeting room environment as presented in the experiments in Chapter 3.
While beamforming could be achieved blindly through automatic array calibration, the two-stage procedure of automatic calibration requiring location coordinates prior to the actual time delay estimation is a somewhat suboptimal approach to beamforming for unknown microphone placements and speaker positions. Furthermore, it should be noted that, this is difficult to accomplish in certain conditions mentioned below:

(i) The array shape calibration proposed in the previous chapter has been shown to work well for closely spaced arrays in office environments. However, it is not suitable for large microphone spacing as the noise becomes increasingly incoherent. This inaccuracy could affect the overall geometrical estimates, in turn reducing the precision of speaker localisation.

(ii) The presence of dominant coherent noise (e.g., noise field generated by a prominent source from a well-defined direction such as hum or cross talk).

(iii) The search space for SRP-PHAT localisation can be so large to cover the whole meeting room. This would be computationally complex and costly to implement.

### 5.2 Blind Estimation Approach

From the MVDR solution in Equation 2.43, the beamformer filter weights are functions of the two parameters which are: the array steering vector $\mathbf{d}$ and the noise coherence matrix $\Gamma_{vv}$.

If the true locations of the microphone and the speaker are known, the array steering vector $\mathbf{d}$ can be easily constructed using Equation 2.28. When these locations are unknown, there are two possible blind beamforming approaches as mentioned in the introduction of this chapter. The first approach to constructing the array steering vector is to employ an automatic calibration method which has been presented in previous chapter to estimate microphone and speaker positions, and then applying Equation 2.28. In the second approach, the delay and gain scaling factors which characterise the incoming signal are estimated from the
received signals directly. This direct estimation of steering vector is referred to as the blind estimation approach.

The direct estimation of array steering vector $\mathbf{d}$ from the recorded array signals consist of gain scaling factor and TDOA computation. These two parameters are estimated from the procedures outlined below:

(1) **Gain Scaling Estimation:**

A two step approach is proposed to estimate the gain scaling factor $a_n$. First, the gain level on each channel due to signal acquisition is normalised. In order to do this, the normalisation factor $\beta_n$ is first calculated and used to normalise the input level on each channel $n$, where:

$$\beta_n = \sqrt{\frac{E_{\text{ref}}}{E_n}}$$

(5.1)

The $E_n$ is calculated as the average of the $M$ lowest energy frames on each channel $n$. The highest energy channel is selected as the reference channel $E_{\text{ref}}$.

The gain scaling factor $a_n$ is then estimated as the ratio of frame energies between the reference channel and each channel, for each time step corresponding to each new delay vector estimate. Assuming a single speaker, calculating the gain factor in this manner should reflect the speech level arriving on each microphone.

(2) **TDOA Estimation:**

The appropriate delay for each microphone $n$, denoted as $\tau_n$, can be determined by finding the TDOA with respect to the reference channel, which corresponds to the peak in GCC function. The TDOA is calculated via cross correlation once every several acoustic frames. To have a robust TDOA estimate, the cross correlation measure is calculated over a large window incorporating several frames, which constitutes a tradeoff between robustness and capturing rapid variations in the TDOA. To aid with the estimation accuracy of the delay, the input signals are pre-emphasised using a two-tap high pass filter ($[1 - 0.95z]$) to attenuate the low frequency band of the signal where
the TDOA estimation is less reliable due to the presence of low frequency noise. To increase the precision of the delays, time domain interpolation is performed to improve the temporal resolution of the peak estimation.

The noise coherence matrix $\Gamma_{vv}$ on the other hand, may be directly estimated using noise samples, either offline or adaptively, or calculated based on an assumed theoretical noise field model. For experiments in this chapter, three different methods to estimate the coherence matrix are outlined below:

(i) Direct Noise Field Estimation: The measured $\Gamma_{vv}$ is estimated using Equation 2.42 and 4.7.

(ii) Spatially Uncorrelated Noise Model: $\Gamma_{vv} = I$, this solution corresponds to the delay-sum beamformer.

(iii) Spherically Isotropic Noise Model: $\Gamma_{vv} = \Gamma_{vv}^{\text{diffuse}}$, this solution corresponds to the superdirective beamformer.

The spherically isotropic (diffuse) noise model requires knowledge of the distances between each pair of microphones. In a scenario when these are unknown \textit{a priori}, these could first be estimated by fitting a diffuse noise model to the measured noise coherence. Based on previous work in Chapter 4, a \textit{blind superdirective beamforming} method is proposed in which the distance between each microphone pair is estimated from the least-squares fitting in Equation 4.5 and 4.6.

5.2.1 Summary of Blind Beamforming Approaches

The beamformer with direct estimation of gain and TDOA for the steering vector operated on spatially uncorrelated noise is termed Blind Delay-Sum beamforming. Similar terms are assigned for beamforming weights obtained using a diffuse noise model and direct estimation from noise samples which are Blind Superdirective and Blind MVDR beamforming respectively.

The automatic calibration approach, in which the array steering vector is obtained after first estimating the microphone and speaker locations, can also
be combined with different type of beamformers. These designs are termed AutoCal+DelaySum, AutoCal+Superdirective, and AutoCal+MVDR for the delay-sum, superdirective, and MVDR beamformer respectively.

Table 5.1 classifies each technique in terms of the method used to estimate the steering vector and the noise coherence matrix.

<table>
<thead>
<tr>
<th>Design</th>
<th>Beamformer</th>
<th>Tech. of array steering estimation</th>
<th>Noise estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blind Estimation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delay-sum</td>
<td>Gain and TDOA estimate</td>
<td>Spatially uncorrelated noise model</td>
<td></td>
</tr>
<tr>
<td>Superdirective</td>
<td>Gain and TDOA estimate</td>
<td>Diffuse model using distance estimate</td>
<td></td>
</tr>
<tr>
<td>MVDR</td>
<td>Gain and TDOA estimate</td>
<td>Estimation from noise samples</td>
<td></td>
</tr>
<tr>
<td><strong>Automatic Calibration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delay-sum</td>
<td>Array shape and source localisation</td>
<td>Spatially uncorrelated noise model</td>
<td></td>
</tr>
<tr>
<td>Superdirective</td>
<td>Array shape and source localisation</td>
<td>Diffuse model using distance estimate</td>
<td></td>
</tr>
<tr>
<td>MVDR</td>
<td>Array shape and source localisation</td>
<td>Estimation from noise samples</td>
<td></td>
</tr>
</tbody>
</table>

5.3 Experimental Results

5.3.1 Database Specifications

Experiments were conducted on a subset of MC-WSJ-AV corpus, which offers an intermediate task between simple digit recognition and large vocabulary conversational speech recognition. The description and specification of the full corpus are detailed in [41]. Only the single-speaker stationary and single-speaker moving sentences which are recorded at the University of Edinburgh (UEDIN) were used and the results are reported on 5000-word task, giving a total of 189 and 190 test sentences respectively.

The layout of Edinburgh meeting room in the MC-WJ-AV corpus used for the experiments is shown in Fig 5.1. Experiments are based on two circular arrays denoted as Array 1 and Array 2, with speaker positions specified by seats, a whiteboard, and a presentation board.

In the single-speaker stationary task, there are six conditions in which the speaker reads sentences from six different positions within the meeting room. The six conditions are: speaker sits at seat 1 (Seat 1), speaker sits at seat 2 (Seat 2), speakers stands near to a presentation board (Presentation), speaker stands near to a whiteboard (Whiteboard), speaker sits at seat 3 (Seat 3), and speaker
Table 5.2: % word error rate for headset, lapel, and single distant microphone

<table>
<thead>
<tr>
<th>Channel</th>
<th>No adaptation</th>
<th>Channel adaptation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Headset</td>
<td>14.8</td>
<td>14.7</td>
</tr>
<tr>
<td>Lapel</td>
<td>26.3</td>
<td>20.7</td>
</tr>
<tr>
<td>SDM</td>
<td>87.6</td>
<td>70.1</td>
</tr>
</tbody>
</table>

sits at seat 4 (Seat 4). In the single-speaker moving task, the speaker is moving between those six locations. In the moving speaker task the speaker walks back and forth between two locations while reading sentences. The six conditions for this task are: speaker is walking from seat 1 to seat 2 (Seat 1 to Seat 2), speaker is walking from seat 2 to presentation board (Seat 2 to Pres.), speaker is walking from presentation board to whiteboard (Pres. to Wbrd.), speaker is walking from whiteboard to seat 3 (Wbrd. to Seat3), speaker is walking from seat 3 to seat 4 (Seat 3 to Seat 4), and speaker is walking from seat 4 to seat 1 (Seat 4 to Seat 1).

The standard WSJCAM0 training set was used to train a basic Gaussian Mixture Model (GMM) based HMM speech recognition system using HTK [89]. The HMMs used 13 MFCCs (including the 0th cepstral coefficient) with their first, second, and third order derivatives. The dictionaries used were generated from that developed for the AMI NIST RT05S system [30], and the standard MIT-Lincoln labs 5k Wall Street Journal trigram language model was used for decoding. To compensate for the channel mismatch between training and testing conditions, the baseline models were adapted for each technique using a static two-pass MLLR adaptation. All results shown include this adaptation. The adaptation data for stationary-speaker and moving-speaker task were taken from their corresponding UEDIN DEV set and consisted of 289 sentences and 251 sentences respectively.

For speech recognition results, every speaker condition %WER in both stationary or moving tasks is calculated using Equation 3.15 and the overall result is similarly obtained from the cumulative totals.

There are two array geometries on which the beamforming algorithms were tested:
5.3. Experimental Results

Figure 5.1: Edinburgh meeting room, according to [41] (measurements in cm).

1. *Circular Array 1.* A fixed 8-element, equally spaced, circular microphone array with a diameter of 20 cm (denoted as Array 1 using microphones 1 to 8 in Figure 5.1), and

2. *Ad-hoc Array.* This geometry also uses 8 microphones elements but selects 4 channels from Circular Array 1 and 4 channels from Circular Array 2 in Figure 5.1. The 4 channels are chosen randomly from each of the array for every ten utterances in both development and evaluation sets. The ad-hoc array geometry simulates randomly placed microphones. For this array, some inter-microphone distances are small, while others are relatively large, as the distance between Circular Array 1 and Circular Array 2 is about 1.5 metres.

5.3.2 Baseline Experiments

To provide a benchmark comparison, speech recognition results generated using headset, lapel, and Single Distant Microphone (SDM) are shown in Table 5.2. Note that for the SDM, microphone 1 of Circular Array 1 was selected.

A first set of baseline experiments on the single stationary speaker task was conducted in order to facilitate direct comparison of results with those reported previously on the MC-WSJ-AV corpus in [41]. For comparison with that work,
Chapter 5. Blind Beamforming

Table 5.3: % word error rate for blind delay-sum beamforming

<table>
<thead>
<tr>
<th>Speaker Location</th>
<th>16 Microphones</th>
<th>8 Microphones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seat 1</td>
<td>43.1</td>
<td>34.7</td>
</tr>
<tr>
<td>Seat 2</td>
<td>31.6</td>
<td>31.6</td>
</tr>
<tr>
<td>Presentation</td>
<td>42.7</td>
<td>59.9</td>
</tr>
<tr>
<td>Whiteboard</td>
<td>43.0</td>
<td>63.8</td>
</tr>
<tr>
<td>Seat 3</td>
<td>25.7</td>
<td>28.0</td>
</tr>
<tr>
<td>Seat 4</td>
<td>35.3</td>
<td>24.9</td>
</tr>
<tr>
<td>All Positions (6)</td>
<td>36.9</td>
<td>40.3</td>
</tr>
<tr>
<td>Seat Positions Only (4)</td>
<td>34.2</td>
<td>30.3</td>
</tr>
</tbody>
</table>

The blind delay-sum beamformer was used, with the addition of a Wiener pre-filter stage having transfer function of \( \frac{P_{xx}(f) - P_{nn}(f)}{P_{xx}(f)} \), where \( P_{xx}(f) \) is the input signal spectrum and \( P_{nn}(f) \) is the input noise spectrum taken from the M lowest energy frames. The results of these experiments are shown in Table 5.3, broken down according to each speaker location, as well as according to whether all 16 (both circular arrays from the corpus) or only 8 (Circular Array 1) microphones were used. The results show that when using all 16 microphones, the blind delay-sum method achieves 36.9%, in direct comparison to the result of 36.8% reported in [41] (achieved without a gain normalisation step).

Looking at the result break-down, the overall results from the delay-sum of 16 microphones achieves a lower WER compared to the delay-sum of 8 microphones. When results are examined by conditions, the presentation and whiteboard contribute to a significant increase in WER, particularly for Circular Array 1. This may be attributed to the longer distances between those two positions and circular array 1, leading to a poorer input SNR due to noise and reverberations. If the presentation and whiteboard locations are removed, the WER using circular microphone array 1 for the four seated positions lowers to 30.3%, compared to 34.2% if using both circular microphone arrays. This again shows that the distance between the source and the array is a key factor in achieving good performance, as the smaller 8 microphone array out-performs the 16-channel array due to its proximity to the speech sources.

Due to these factors, to yield more straightforward comparison between blind
Table 5.4: % word error rates for stationary speaker task using a circular array in various speaker positions.

<table>
<thead>
<tr>
<th>Speaker Location</th>
<th>Blind Delay-Sum</th>
<th>Blind Superdirective</th>
<th>Blind MVDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seat 1</td>
<td>37.1</td>
<td>33.8</td>
<td>36.7</td>
</tr>
<tr>
<td>Seat 2</td>
<td>32.4</td>
<td>27.4</td>
<td>27.6</td>
</tr>
<tr>
<td>Seat 3</td>
<td>31.1</td>
<td>24.1</td>
<td>28.8</td>
</tr>
<tr>
<td>Seat 4</td>
<td>30.6</td>
<td>24.2</td>
<td>28.4</td>
</tr>
<tr>
<td>Overall</td>
<td>33.0</td>
<td>27.8</td>
<td>30.6</td>
</tr>
</tbody>
</table>

Table 5.5: % word error rates for stationary speaker task using a circular array in various speaker positions.

<table>
<thead>
<tr>
<th>Speaker Location</th>
<th>AutoCal+Delay-Sum</th>
<th>AutoCal+Superdirective</th>
<th>AutoCal+MVDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seat 1</td>
<td>39.5</td>
<td>35.9</td>
<td>39.6</td>
</tr>
<tr>
<td>Seat 2</td>
<td>34.2</td>
<td>28.9</td>
<td>33.7</td>
</tr>
<tr>
<td>Seat 3</td>
<td>29.6</td>
<td>26.3</td>
<td>28.8</td>
</tr>
<tr>
<td>Seat 4</td>
<td>32.2</td>
<td>26.3</td>
<td>31.5</td>
</tr>
<tr>
<td>Overall</td>
<td>34.2</td>
<td>29.7</td>
<td>33.8</td>
</tr>
</tbody>
</table>

beamforming approaches, the subsequent stationary speaker experiments focus on the use of the single 8-microphone circular array for the four seated locations.

5.3.3 Comparison of Blind Approaches

As mentioned above, results in this section are only reported for beamforming techniques using Circular Array 1 (8 microphones) for Seat 1, Seat 2, Seat 3, and Seat 4 - a total of 128 test utterances. The stationary speaker results for the three different blind beamforming methods are presented in Table 5.4, and those for the methods based on array calibration are presented in Table 5.5. Note that in order to directly compare the beamforming approaches in the subsequent experiments, the Wiener pre-filter from the baseline experiments above is no longer applied. This has the effect of increasing the WER for blind delay-sum beamforming to 33.0% (in Table 5.4) from 30.3% (in Table 5.3).

Results for the moving speaker task for the three different blind beamforming
Table 5.6: % word error rates for moving speaker task using a circular array in various speaker movements.

<table>
<thead>
<tr>
<th>Speaker Movement</th>
<th>Blind Delay-Sum</th>
<th>Blind Superdirective</th>
<th>Blind MVDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seat 1 to Seat 2</td>
<td>38.4</td>
<td>33.9</td>
<td>36.0</td>
</tr>
<tr>
<td>Seat 2 to Pres.</td>
<td>57.3</td>
<td>45.6</td>
<td>45.3</td>
</tr>
<tr>
<td>Wbrd. to Seat 3</td>
<td>51.2</td>
<td>49.8</td>
<td>52.1</td>
</tr>
<tr>
<td>Seat 3 to Seat 4</td>
<td>49.8</td>
<td>44.3</td>
<td>45.5</td>
</tr>
<tr>
<td>Overall</td>
<td>49.1</td>
<td>43.2</td>
<td>44.5</td>
</tr>
</tbody>
</table>

Table 5.7: % word error rates for stationary speaker task using an ad-hoc array in various speaker positions.

<table>
<thead>
<tr>
<th>Speaker Location</th>
<th>Blind Delay-Sum</th>
<th>Blind Superdirective</th>
<th>Blind MVDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seat 1</td>
<td>61.9</td>
<td>54.9</td>
<td>58.3</td>
</tr>
<tr>
<td>Seat 2</td>
<td>41.7</td>
<td>40.7</td>
<td>43.8</td>
</tr>
<tr>
<td>Seat 3</td>
<td>35.4</td>
<td>34.6</td>
<td>38.7</td>
</tr>
<tr>
<td>Seat 4</td>
<td>52.1</td>
<td>48.8</td>
<td>50.5</td>
</tr>
<tr>
<td>Overall</td>
<td>47.9</td>
<td>44.8</td>
<td>48.0</td>
</tr>
</tbody>
</table>

methods are presented in Table 5.6. For these results, only Seat 1 to Seat 2, Seat 2 to Pres., Wbrd. to Seat 3, and Seat 3 to Seat 4 conditions are reported, which is also a total of 128 test utterances. Note that the methods based on array calibration are not presented for the moving speaker task, as this would necessitate a more complex speaker tracking algorithm, rather than the SRP-PHAT localisation method. This highlights the relative simplicity of the direct estimation blind beamforming approaches, making them favourable for a single speaker situation.

Similarly, results using the ad-hoc array (defined in Section 5.3.1) for the stationary and moving speaker tasks for the three different blind beamforming methods are presented in Tables 5.7 and 5.8, respectively.
5.4. Discussion

(a) Directivity patterns of a circular array with delay-sum beamformer weights. The beamformer is directed towards a speaker Seat 1 at 1 metre.

(b) Directivity patterns of an array of 4 microphones (mic. 2, 4, 6, and 8) with delay-sum beamformer weights. The beamformer is directed towards Seat 1 at 1 metre.

(c) Directivity patterns of an ad-hoc array of 8 microphones with delay-sum beamformer weights. Microphones for ad-hoc array are chosen from microphone 2, 4, 6, 8, 10, 12, 14, and 16. The beamformer is directed towards Seat 1 at 1 metre.

Figure 5.2: Directivity patterns plots.
Table 5.8: % word error rates for moving speaker task using ad-hoc array in various speaker movements.

<table>
<thead>
<tr>
<th>Speaker Movement</th>
<th>Blind Delay-Sum</th>
<th>Blind Superdirective</th>
<th>Blind MVDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seat 1 to Seat 2</td>
<td>48.2</td>
<td>47.9</td>
<td>50.8</td>
</tr>
<tr>
<td>Seat 2 to Pres.</td>
<td>46.0</td>
<td>45.6</td>
<td>49.2</td>
</tr>
<tr>
<td>Wbrd. to Seat 3</td>
<td>49.4</td>
<td>47.9</td>
<td>51.5</td>
</tr>
<tr>
<td>Seat 3 to Seat 4</td>
<td>53.8</td>
<td>53.8</td>
<td>54.9</td>
</tr>
<tr>
<td>Overall</td>
<td>49.2</td>
<td>48.7</td>
<td>51.5</td>
</tr>
</tbody>
</table>

5.4 Discussion

For the stationary speaker task, and using an appropriately close cluster of 8 microphones, Table 5.4 shows that blind superdirective beamforming achieves the best performance amongst the investigated methods. This is followed by blind MVDR with slightly higher WER. The absolute WER improvement between superdirective and delay-sum is 5.2%, and MVDR and delay-sum is 2.4%. A similar trend is observed in the moving speaker task, in which blind superdirective and MVDR are superior to delay-sum, achieving performance of 5.9% higher for superdirective and 4.6% for MVDR. When both superdirective and MVDR are compared, superdirective beamforming achieves slightly better performance than MVDR in both stationary and moving tasks for a closely-spaced array located near to the speaker. Similar results, albeit degraded by 1-3%, are achieved using the methods based on array calibration, as shown in Table 5.5. This degradation is likely attributable to steering errors caused by imprecise microphone positions estimated in the calibration step [47].

The improvement of superdirective and MVDR over delay-sum is attributed to the high degree of noise correlation that exists at low frequencies when the distance between two microphones is small [12]. The spatial coherence may be captured either by calculating the coherence matrix directly from noise samples, as in the case of MVDR, or modelled using a sinc function (Equation 2.14), as in the case of superdirective. In the current experiments, the use of the diffuse noise field model yields the best performance.
5.4. Discussion

From Table 5.7, using the ad-hoc array geometry instead of the circular array for the stationary speaker task leads to a WER increase from 33.0% to 47.9% for the blind delay-sum beamformer, that is, a degradation of 14.9% of speech recognition performance. Similarly, the performance of blind superdirective and MVDR degrade when using the ad-hoc array rather than the circular array, with WER increases of 17.0% and 17.4%, respectively. As a result, for the ad-hoc array the performance of blind delay-sum, superdirective, and MVDR are similar. Superdirective still achieves the best results, although the difference in performance compared to delay-sum is only 3.1% absolute. For the moving speaker task, all techniques achieve the same level of performance.

As discussed above, for closely-spaced arrays such as the circular array, it is more effective for the noise coherence matrix $\Gamma_{vv}$ to be estimated from their theoretical model or to be estimated directly from noise samples. However, as the distance between microphones increases, as is the case in the current ad-hoc array geometry, the noise becomes less correlated between channels. This means that the measured and modelled coherence for the MVDR and superdirective methods will approach the incoherent model assumed in the delay-sum method, and so their performance is expected to converge.

In the stationary speaker case, when using the ad-hoc array instead of the circular array, the overall speech recognition performance is significantly reduced for all beamforming techniques. This is the case even though the number of microphones used is the same. This degradation is attributable to two main factors. First, a microphone array benefits from basic design of its geometry for optimal effectiveness in different scenarios. In this instance, the ad-hoc array has inappropriate inter-element spacings leading to spatial aliasing effects in its directivity pattern. Spatial aliasing causes the appearances of grating lobes in directions other than the look direction. The extent of this effect in the current context is demonstrated in the directivity patterns plotted in Figure 5.2(a) (8-element circular array), Figure 5.2(b) (4 elements from a single circle array within the ad-hoc array) and Figure 5.2(c) (8-element ad-hoc array).

Apart from this spatial aliasing problem, a second cause of degraded performance is in the fact that some microphones are located further from the desired
speaker, in turn leading to a naturally lower input SNR than for the first circular array, in which all elements are close to the seated locations. This effect is shown in the fact that the ad-hoc array yields better performance than the circular one for transitions involving the whiteboard or presentation areas in the moving speaker task, as the speaker moves closer to the elements from the second circular array that are included in the ad-hoc array.

These results suggest that for truly ad-hoc arrays it may often prove more effective to first select a subset of microphones that are relatively close to each other and the desired speaker, rather than beamforming using all microphones. This will minimise potential spatial aliasing effects, ensure the best possible input SNR prior to enhancement, and allow greater noise reduction based on its coherence across channels.

5.5 Summary

In this chapter, different blind methods for achieving microphone array beamforming were investigated in the context of a speech recognition system. Single source beamforming methods are generally characterised by two terms - the steering vector and the noise coherence (or correlation) matrix. Two principal techniques of estimating the steering vector were investigated: blind delay estimation and automatic array calibration methods. Similarly, the noise environment can be either estimated directly from noise samples, or else approximated using theoretical models. These different techniques to estimate both the steering vector and noise coherence matrix were combined to form different methods to beamform from unknown array geometries.

Experiments were conducted on the single speaker portions of the MC-WSJ-AV corpus using two different array geometries - an 8-element circular array in the middle of a meeting table, and an 8-element ‘ad-hoc’ array, simulated by combining random elements from across the 16 channels available in the corpus. Overall results confirm that beamforming methods based on delay-sum are suboptimal for closely spaced microphone arrays in an office environment, as the noise coherence between channels can be significant. In such conditions, a significant
improvement is gained using blind superdirective or MVDR approaches to better model the noise conditions.

While methods based on a first array calibration step were effective, in the experiments it was shown that more precise blind beamforming can be achieved from delay estimates directly, without the added complexity of first estimating the microphone positions. Array calibration based approaches are however expected to be preferable when trying to track, separate and recognise multiple speakers with potential overlap, as shown in Chapter 4. It is also noted that the blind superdirective beamforming method is reliant on the first step from the calibration routine, or some other method of estimating the inter-microphone distances.

As microphone spacings increase, as in the ad-hoc array experiments, the noise coherence between channels becomes less significant and, as expected, the superdirective and MVDR approaches only give minor improvements over the less complex blind delay-sum method. Some of the results presented also indicated that best performance was achieved when microphones are placed closest to the speaker, as this assures a higher input SNR even prior to any enhancement. Thus including more channels in the array may not necessarily lead to best performance when these channels have a low input SNR. A further significant factor with the ad-hoc array investigated in these experiment was the fact that spatial aliasing may arise when microphone spacings increase.

Based on these findings, the next chapter will investigate clustering approach for robust beamforming from ad-hoc microphone arrays. By first performing a classification of available microphones into closely-spaced clusters located close to the desired speech sources, blind superdirective or MVDR methods may be applied to achieve robust noise reduction through spatial filtering, and hence good speech recognition performance. For truly ad-hoc deployments, this follows the sensor network philosophy in which a subset of all available nodes is used to perform locally relevant operations. Further work is also continuing to record a data set with ad-hoc microphone placements.
Chapter 6

Clustered Approach for Ad-hoc Microphone Arrays

6.1 Introduction

In dealing with microphone array beamforming for ad-hoc array geometries, two approaches have been proposed in Chapter 5. These are automatic calibration and blind estimation. Beamforming using automatic calibration is based on the geometrical representation of the arrays. While achieving accurate calibration, there can be a geometrical constraint such as when the array is large with wide inter-microphone spacing. Blind beamforming on the other hand relies on the direct calculation of the relative time delay between microphones, thus it works for any array geometry. Both approaches however give a similar performance in terms of speech recognition accuracy using various beamforming techniques for small ad-hoc arrays.

It is easy then to conclude that the solution to the problem of beamforming for ad-hoc array geometry is to do the beamforming blindly, however from experiments presented in the previous chapter, simply using all microphones for beamforming does not always yield the highest performance. While the achievable array gain generally increases with the number of elements, it is commonly assumed that microphones are physically identical and are located spatially close to each other to have similar acoustic conditions.
When a microphone is arbitrarily placed, the signal acquired by the transducer depends on its characteristics such as gain and directional response and the acoustic conditions of the room involving reverberation and the presence of localised noise sources. This means that some microphones would be better discarded due to their poor input SNR and signal quality. A particular consideration is that large inter-microphone spacing may lead to an erroneous time difference of arrival computation, effectively causing delay inaccuracies in the steering vector of the beamformer. It is also undesirable to have spatial aliasing effects in the beamformer’s directivity pattern as shown in previous chapter. Therefore, it is hypothesised that using a cluster of microphones (ie, a sub-array), closely located both to each other and to the desired speech source may in fact provide a more robust speech enhancement than using the full array. In ad-hoc situations, the lack of prior knowledge of microphone and speaker locations means that the clustering of microphones and the selection of clusters must be done blindly.

This chapter proposes a novel clustered approach for beamforming from ad-hoc microphone arrays. As mentioned above, for ad-hoc microphone arrangement, a particular issue is that large inter-microphone spacings may lead to erroneous time difference of arrival computation, effectively causing delay inaccuracies in the steering vector of the beamformer. Hence, the theoretical justifications of using a subset of microphones instead of the full array as a more robust approach are analysed in Section 6.1 below, motivating methods for selecting a cluster of microphones that are both close to each other, as well as to the speaker. With this motivation, first, an inter-microphone proximity measure is proposed based on the Magnitude Squared Coherence (MSC) function during noise periods. Two different algorithms for forming clusters based on this measure are proposed, followed by a method for ranking the formed clusters according to their proximity to the desired speaker. Using the ranked clusters, beamforming may then be achieved by applying the blind methods from Table 5.1 to either the closest cluster, or else employing a weighted combination of all clusters.
6.2 Theoretical Justification

The array gain \( G \) measures the SNR improvement between one sensor and the output of the whole array. This is given by Equation 2.51 where the signal model and \( \mathbf{d} \) have been given in Equation 2.24 and 2.28 respectively.

For delay-sum beamformer assuming noise is uncorrelated from sensor to sensor (i.e. \( \Gamma_{vv} = \mathbf{I} \) thus \( \mathbf{w} = \frac{1}{N} \mathbf{d} \) in Equation 2.43) and the steering delays are matched to the wave’s direction of propagation, the array gain can be simplified to:

\[
G = \left| \frac{1}{\sum_{i=1}^{N} |w_i|^2} \right| = N
\]  

(6.1)

where \( N \) is the number of sensors in the array.

Adding one further microphone to the array, the array gain which consists of previously \( N \) microphones can be written as:

\[
G = \sqrt{\sum_{i=1}^{N+1} \frac{a_i}{N+1} + \frac{b}{N+1} e^{j2\pi \Delta \tau}}^2
\]  

(6.2)

where \( \Delta \tau \) is the small variation in delay estimates due to mismatches of the steering delays to the wave’s direction of propagation. Assuming each microphone gain has \( \beta = a_i = 1 \ \forall i \) and \( \Delta \tau = 0 \) for including a microphone with matched delays, leading to in-phase addition of the complex frequency signals, the maximum gain achieved is \( G_{hi} = N + 1 \). If the addition of a microphone causes phase difference of \( \pi \) radians (i.e. antiphase addition), the lower bound of array gain for including this one microphone is equal to:

\[
G_{lo} = \left| \frac{N}{N+1} - \frac{1}{N+1} \right|^2 = \frac{(N - 1)^2}{N + 1}
\]  

(6.3)

where the complex exponential in the numerator of Equation 6.2 is equal to \( e^{-j\pi} = -1 \) due to the \( \theta - \pi \) phase shift from the sum of other microphones signals with matched delays (in which \( \theta \) is assumed equal to 0).

In this case of including an additional one microphone, the array gain will vary between \( G_{lo} \leq N \leq G_{hi} \). As a concrete example, for \( N = 9 \), the theoretical gain from a single additional microphone with incorrect phase alignment may vary from \( G = 6.4 \) to \( G = 10 \). It is then a robust strategy to exclude a microphone with large delay errors since the overall gain may be less than \( N \).
Figure 6.1: Box plot of the TDOA values applied to 8-element microphone arrays. The lines are drawn for the lower quartile, median, and upper quartile values. Whiskers extend to one standard deviation above and below the mean (circle) of the data. The expected ground truth of TDOA values are shown in asterisk. In the x axis, TDOA\textsubscript{21} corresponds to the TDOA between the second closest and the closest microphone in the array, similarly, TDOA\textsubscript{31} corresponds to the TDOA between the third closest and the closest microphone in the array, and so on.

Figure 6.1 shows TDOA values calculated from the speech frames between a 8-element microphone array. The TDOAs are calculated between a reference microphone to 3 closely located microphones (within 20cm of distance) and other 4 microphones of about 2m from the reference microphone. The empirical observation shows that TDOAs are accurate between closely spaced sensors, with accuracy decreasing with greater distance between the pair used to calculate TDOA (e.g. TDOA\textsubscript{51}, TDOA\textsubscript{61}, TDOA\textsubscript{71}, and TDOA\textsubscript{81}) due to the large distance between the reference microphone and the 5\textsuperscript{th}, 6\textsuperscript{th}, 7\textsuperscript{th}, 8\textsuperscript{th} closest microphone to the speaker respectively.

Consider the effect of variations in microphone gain $a$, $0 < a < 1$ in Equation 2.24. From Equation 2.29, $a$ relates to the distance from source to the microphone in an ideal propagation model. Assuming all microphones in a cluster have matched delays, the overall gain is the sum over gain factors for a cluster,

$$G = \left| \frac{\sum_{i=1}^{N} a_i}{\sum_{i=1}^{N} |w_i|^2} \right|^2 = \frac{1}{N^2} \left| \sum_{i=1}^{N} a_i \right|^2 = \Lambda N$$

where $\Lambda = \frac{1}{N^2} \left| \sum_{i=1}^{N} a_i \right|^2$, in which $\Lambda$ has value between 0 and 1. From Equation 6.4, the total gain of a cluster depend on the two variables, the gain factor $a$ and the size of cluster $N$.

Assume an array of $N$ microphones consists of two clusters having $N_1$ and
\( N_2 \) number of microphones respectively (i.e. \( N = N_1 + N_2 \)), the ratio of cluster gains,

\[
\frac{G_1}{G_2} = \frac{N_1}{N_2} \frac{1}{N_1} \left( \sum_{i=1}^{N_1} \frac{d_{ref}}{\| s - m_1 \|} \right)^2 \frac{1}{N_2} \left( \sum_{i=1}^{N_2} \frac{d_{ref}}{\| s - m_2 \|} \right)^2
\]

(6.5)

where the sensor gain \( a \) has been replaced with the ratio between distance from source to the reference microphone \( d_{ref} \) and the distance from source to each microphone in the cluster. Here, the \( d_{ref} \) is the distance from source to the closest microphone in the array which is the same for both clusters. From Equation 6.5, the ratio of cluster gains depends in general on the size of cluster and inversely on the average distance of that cluster from the sound source.

From this analysis, robust gain will be achieved when:

1. Only elements with accurate inter-microphone delays estimates are included, and
2. The gain factor \( a \) is close to 1 which means that the microphones are closer to the source.

### 6.3 Inter-Microphone Proximity Measure

The MSC between two microphone signals \( i \) and \( j \) at discrete frequency \( f \), \( C_{ij}(f) \) is calculated in the following manner:

\[
C_{ij}(f) \triangleq \frac{|\Phi_{v_iv_j}(f)|^2}{\Phi_{v_i}(f)\Phi_{v_j}(f)}
\]

(6.6)

where the auto- and cross-power spectral densities are estimated as in Equation 4.7.

Environments such as offices or meeting rooms are usually considered to represent diffuse noise fields. The MSC function between two microphones in a diffuse noise field can be modelled as [12]:

\[
C_{ij}^{\text{diff}}(f) = \sin^2 \left( \frac{2\pi f d_{ij}}{c} \right)
\]

(6.7)

According to this model, the noise coherence between two microphones depends principally on the distance \( d_{ij} \) between them. The first minimum of this MSC
Figure 6.2: Theoretical magnitude squared coherence as a function of distance between two microphones. The sampling frequency is 16 kHz.

function occurs at:

\[ f_m(i, j) = \frac{c}{2d_{ij}} \]  

(6.8)

and beyond this frequency the coherence approaches zero.

This dependence of the diffuse noise coherence on the distance can be used to indicate how close two microphones are, since closely-spaced microphones will have wider main lobes in the coherence function compared to distantly-spaced pairs as shown in Figure 6.2. To give a measure of overall coherence between microphones, and hence a measure of their proximity, the MSC may be integrated across frequencies:

\[ T_{ij}^{MSC} = \sum_{0}^{f_{max}} C_{ij}(f) \]  

(6.9)

where the summation range is limited by \( f_{max} \) to improve robustness. Typically, this may be set to be \( f_m(i, j) \) from Equation 6.8, as the measured coherence function often varies significantly from the theoretical model for frequencies much beyond the main lobe.

6.3.1 Rule-based Clustering

In order to cluster microphones, the measure in Equation 6.9 may be compared to some threshold value to determine if two microphones are sufficiently close to each
other. A threshold value can be computed to correspond to a desired distance, \( d_\varepsilon \), by using the theoretical coherence model from Equation 6.7 and summing up to a threshold frequency \( f_\varepsilon = c/2d_\varepsilon \) (corresponding to its first minimum):

\[
T_\varepsilon = \sum_{l=0}^{f_\varepsilon} \sin^2 \left( \frac{2\pi f_\varepsilon d_\varepsilon}{c} \right)
\]  

(6.10)

The measured value for \( T_{MSC}^{ij} \) may then be compared to this intra-cluster threshold \( T_\varepsilon \). If \( T_{MSC}^{ij} \geq T_\varepsilon \), then microphones \( i \) and \( j \) are grouped in the same cluster, otherwise they belong to separate clusters. This conservative binary classification is evaluated over all microphone pairs to form an initial set of clusters in which all microphones are within the specified distance of all others in the cluster. A subsequent merging pass then combines clusters for which at least one inter-cluster microphone pair is within a more restrictive distance. Without this second pass, the method would only capture clusters with maximum extent defined by the distance threshold - the merging step allows larger clusters to be formed from these if continuity exists at the boundaries. The proposed clustering algorithm is as follows:

1. Assign \( b_{ij} = 1 \) if \( T_{MSC}^{ij} \geq T_\varepsilon \), for \( i, j = 1, \ldots, N \).
2. Compute \( B_i = \sum_{j=1}^{N} b_{ij} \), for \( i = 1, \ldots, N \).
3. Select the microphone belonging to the most pairs as the centre microphone of the first cluster, i.e. \( \hat{m}_1 = \arg \max_i B_i \).
4. Form cluster 1, \( Q_1 \) with \( Q_1 = \{ j | b_{\hat{m}_1j} = 1 \} \).
5. Remove microphones belonging to cluster 1 from consideration, then repeat the above steps to form clusters \( k = 2 : K \) until all microphones have been assigned a cluster.
6. Once the set of initial clusters \( Q_{1:K} \) is formed, a merging pass is conducted. Two clusters are merged if \( T_{MSC}^{ij} \geq T_\kappa \), where microphone \( i \) belongs to one of the clusters and \( j \) belongs to another, and \( T_\kappa \) is an inter-cluster threshold calculated using a more restrictive distance criteria \( d_\kappa \) in Equation 6.10.
7. In the case the above steps result in the formation of any single-element clusters, these may be merged with the closest cluster if a relaxed inter-cluster threshold is satisfied.

6.3.2 Spectral Clustering

As an alternative to the rule-based clustering algorithm explained above, the use of spectral clustering is also investigated. Spectral clustering finds group structure in data by using the spectrum of a similarity matrix. The algorithm is based on spectral graph theory and has been widely used for pattern recognition [55, 72, 91]. Here spectral clustering is applied on the coherence measure to automatically group microphones in the spatial domain without need for hard decision rules.

The key to the partitioning is the construction of the similarity matrix as a weighted adjacency matrix $S$ modeling the neighborhood relationship between data points. To cluster microphones using the MSC measure, this matrix may be defined as:

$$s_{ij} = \exp\left(-\alpha\left(\frac{1}{L_{f_e}}T_{MSC}^{ij} - 1\right)^2\right) \tag{6.11}$$

where $L_{f_e}$ is the DFT length up to frequency $f_e$, used to normalise the sum of MSC values $T_{MSC}^{ij}$, and $\alpha$ is the scaling parameter in the Gaussian filtering function.

The algorithm of spectral clustering investigated in this chapter follows [72, 84]:

1. Construct matrix $S = (s_{ij})_{i,j=1,\ldots,N}$ defined in Equation 6.11.

2. Compute $D = \text{diag}(d_1, \ldots, d_i)$, where $d_i$ is defined as:

$$d_i = \sum_{j=1}^{N} s_{ij} \tag{6.12}$$

3. Solve the generalised eigenproblem $(D - S)v = \lambda Dv$ for the first $k$ eigenvectors, where $k$ is the number of clusters.

4. Form matrix $V \in \mathbb{R}^{N \times k}$ containing the vectors $v_1, \ldots, v_k$ as columns.

5. Partition matrix $V$ into vector $y_i \in \mathbb{R}^k$ for $i = 1, \ldots, N$ which correspond to the $i^{th}$ row of $V$. 
6.4. Ranking Clusters

6. Cluster $y_1, \cdots, y_N$ into $k$ clusters with k-means algorithm in $\mathbb{R}^k$ into clusters $A_1, \cdots, A_k$.

7. Microphones will be clustered into $Q_1, \cdots, Q_k$ with $Q_i = \{j | y_j \in A_i\}$.

The eigengap which indicates the stability of the eigenstructure can reveal the number of clusters \cite{55}. In the ideal case of $k$ completely disconnected clusters, the eigenvalue 0 has multiplicity of $k$ and there will be a gap to the $(k+1)^{th}$ eigenvalue which is greater than zero. Thus the sudden increase of the $k^{th}$ eigengap $\xi_k$,

$$\xi_k = |\lambda_{k+1} - \lambda_k|$$ (6.13)

where $\lambda_k$ is the $k^{th}$ smallest eigenvalue, may indicate the true number of clusters.

Similar to step 7 in the rule-based clustering algorithm, in the case of formation of any single-element clusters, these may be merged with the closest cluster if a relaxed inter-cluster threshold is satisfied.

6.4 Ranking Clusters

Once microphone clusters have been formed according to one of the above methods, it is necessary to somehow select which cluster, or clusters, will be used for beamforming. While a range of signal quality criteria may be used to achieve this more generally, this section proposes one method to achieve this based on the assumption that the best clusters will be those located closest to the speaker of interest.

Assuming a known period of speech from a single person, the delay in receiving a sound wave between clusters indicates their relative distance to that speaker. The detailed steps to rank clusters based on their proximity to the speaker are outlined below. The algorithm considers both the proximity of the closest microphone within each cluster (using the TDOAs between a reference microphone from each cluster), as well as the spatial extent of the cluster (using a measure of the spread of TDOAs within each cluster).
1. Find the closest microphone to the speaker within each cluster and set it as reference $m_k$. To do this, for cluster $k$, choose an initial arbitrary reference microphone $m_k'$ and calculate $\tau(i, m_k')$ for each microphone $i$ in the cluster during speech-only segments. Update the reference microphone for the cluster to be the closest microphone by selecting the one having the minimum TDOA, ie $m_k = \arg \min_i \tau(i, m_k')$.

2. As a measure of cluster spread, calculate the mid-range TDOA offset for each cluster $\delta_k$ relative to its reference microphone. To do this, for cluster $k$, calculate the TDOA of each other microphone with respect to the reference microphone selected in the previous step, ie $\tau(i, m_k)$ for each microphone $i$ in the cluster. Set the mid-range TDOA offset for the cluster to be half of the maximum TDOA, ie $\delta_k = \frac{1}{2} \max_i \tau(i, m_k)$.

3. Find the reference cluster $c_{\text{ref}}$ as the one with its reference microphone closest to the speaker. To do this, first choose an arbitrary reference cluster $k_r$ and calculate the set of TDOAs between its reference microphone and the reference in all clusters $k$, $\tau(m_k, m_{k_r})$. Update the reference cluster to be the one which has the minimum TDOA, ie $c_{\text{ref}} = \arg \min_k \tau(m_k, m_{k_r})$.

4. Form the final proximity score $D_k$ for each cluster by compensating the inter-cluster TDOAs with the cluster mid-range offsets. Given the set of mid-range offsets for all clusters, $\delta_k$, and the set of TDOAs with respect to the reference cluster $c_{\text{ref}}$, $D_k = \tau(m_k, m_{c_{\text{ref}}}) + \delta_k - \delta_{c_{\text{ref}}}$.

The clusters may then be ranked according to their proximity to the speaker according to the score $D_k$. Note that it is possible that this score may be negative, indicating that the reference cluster from step 3 did not turn out to be the closest cluster when considering the mid-range offsets. Considering these mid-range TDOA offsets is a means to compensate for the differing spatial extents of clusters. For instance, while some clusters may have a reference microphone that
is close to the speaker, they may also be large clusters with other microphones that are quite far from the speaker.

6.5 Clustered Beamforming

Following clustering of microphones and the subsequent ranking of clusters according to their proximity to the speaker, blind beamforming may be performed for speech enhancement and recognition. This section presents two methods for beamforming using the clustering information: beamforming using the closest cluster only, and forming a weighted combination over multiple clusters.

6.5.1 Closest Cluster Beamforming

In the spatially distributed microphones scenario, a speaker is usually relatively close to one or subset of microphones in the same time. The simple strategy for speech enhancement in this situation is to choose the signal from microphones which are closest to the speaker or ones which have the best SNR. Assuming identical microphone gains and no obstructions in the line of sight from the speaker to microphones, the microphones which have the highest SNR will be the ones which are closest to the speaker.

6.5.2 Weighted Cluster Combination Beamforming

While the closest cluster may have the best SNR, it is hypothesised that contributions from every cluster may in fact improve the overall performance. In this chapter, delay-sum beamforming is used to combine the beamformed signals of each clusters. To calculate the overall combination, each cluster output is phase-aligned with an estimated fixed delay before a weighted summation.

Given the \( y_k(f) \) is the beamforming output of cluster \( k \) as defined in Equation 2.34, the weighted combination of cluster beamforming output is obtained from

\[
  z(f) = \sum_{k=1}^{K} w'_k H(f) y_k(f) \tag{6.14}
\]
The weight $w'_k(f)$ is defined as

$$w'_k(f) = \alpha_k e^{-j2\pi f \tau_k} \quad (6.15)$$

where $\alpha_k$ represent the contribution of each cluster’s $k$ and $\tau_k$ is the relative delay of each cluster’s output signal with the respect to the reference cluster.

### 6.6 Experimental Setup

Experiments were conducted in a meeting room of size 5.3m x 4.4m x 2.7m, as shown in Figure 6.3. The main sources of noise were a PC, laptop, a projector, and air conditioning. To experiment with different ad-hoc array geometries, microphones were mounted in varying positions on two cork boards placed on top of the meeting table. A total of 8 microphones (AKG C417 omnidirectional condenser microphones) were used for each ad-hoc geometry. The microphones were recorded using a MOTU 8pre audio interface and SONAR 8 software, allowing simultaneous, fully synchronised playback and recording of multiple audio channels.

Evaluating methods to deal with ad-hoc microphone placements clearly requires a trade-off between the desire to test as many configurations as possible, and practicalities of conducting and analysing a large body of experiments. To make some coherent conclusions from this initial investigation into microphone clustering, four data sets were recorded to investigate algorithm behaviour in different circumstances, as described in the following sub-sections. The recordings were constrained to a typical meeting room deployment, with 8 microphones placed in different configurations on a meeting table. While endeavouring to explore a variety of configurations in the spirit of ad-hoc situations, scenarios focus on those which highlight the potential benefits and limitations of clustering.

#### 6.6.1 Data Set A

The first data set was collected to evaluate the proposed microphone clustering algorithm and consisted of noise recordings from 20 different ad-hoc array geometries. The 8 microphones were placed within an area of approximately 1m x
Experimental Setup

While constrained, this area serves to test the algorithm behaviour around the range of the inter-cluster distance parameters designed in the rule-based algorithm. Within those recordings, microphones were positioned on the meeting table to reflect variants of ad-hoc positioning into clearly separated cluster of microphones, closely spaced clusters, or just a single large cluster.

6.6.2 Data Set B

The second data set was collected to evaluate the clustered approach for blind beamforming from ad-hoc arrays. The microphones and speakers were configured such that there were clearly separated microphone subsets with a speaker positioned relatively closer to one of these. To achieve this, two clusters with an equal number of microphones were placed near opposite edges of the table, as shown in Figure 6.3. Three different speaker locations were used: Position S1 where the speaker is facing all microphones but is closer to one cluster than the other, Position S2 where the speaker faces only one of the clusters, and Position S3 where the speaker faces both clusters at varying angles. A total of 30 utterances from the WSJCAM0 evaluation set [59] were recorded by playing the clean speech files through a loudspeaker at the various speaker positions. To simulate random microphone placement within these constraints, the microphone positions were rearranged for every 10 recorded sentences.

To simulate the effect of localised noise source such as a competing speaker for a given ad-hoc configuration, babble noise was played during utterance recordings. The babble noise was taken from NOISEX database [83] and the volume was set to be approximately 10dB lower than the main speaker. Three different noise source positions were used in this experiment: at Position N1 and N2 where the noise source is in close proximity to one of the clusters, and at Position N3 where the noise source is located with large distance to any clusters. For the completion of experiment, the recording session is also performed without the localised noise. The illustration of the ad-hoc array setup for Data Set B is shown in Figure 6.3.
Figure 6.3: Meeting room used in Data Sets B and D (measurements in cm).

### 6.6.3 Data Set C

The third data set serves the same purpose as Data Set B. However in this set of experiments, the number of clusters and speaking orientations were configured differently. The illustration of the ad-hoc array setup for Data Set C is shown in Figure 6.4. There are 3 clusters of microphones with 2 of these containing 3 microphones and 1 cluster containing 2 microphones. The two clusters with 3 microphones were placed near opposite edges of the edge of the table, and the cluster with 2 microphones was placed on the right edge of the table. The speaker at Position S1 is facing all microphones but is relatively closer to one cluster than the others, the speaker at Position S2 is facing all microphones from different angles and close to two of the clusters, and the speaker at Position S3 is also facing all microphones with somewhat equally large distances to every cluster. Four localised noise conditions were positioned in the same manner as in Data Set B.
6.7 Experimental Results

Figure 6.4: Meeting room used in Data Set C (measurements in cm).

6.6.4 Data Set D

This data set was recorded for the purpose of evaluating speech recognition performance. The ad-hoc array arrangement is similar to that described in Data Set B (Figure 6.3) with three speaker positions and no localised noise sources. The data consists of 182 utterances from the evaluation set of the WSJCAM0 corpus for each speaker orientation. For every 10 recorded utterances, microphone positions were rearranged to simulate random placements.

6.7 Experimental Results

6.7.1 Microphone Clustering Evaluation

Compared to classification, clustering can be difficult to objectively evaluate, as often there is no correct grouping that can be considered as ground-truth. To evaluate use of the noise coherence feature, the results of the proposed automatic cluster algorithm based on noise recordings were therefore compared to sub-arrays formed by applying the same algorithm to ground-truth distances between known microphone positions. The comparison is illustrated for three ad-hoc geometries
in Figure 6.5. For the rule-based clustering algorithm, the intra-cluster distance threshold \( d_e \) and inter-cluster distance threshold \( d_k \) are set to be 30cm and 20cm respectively. Single element clusters are merged to the nearest cluster if they satisfy a relaxed threshold of 50cm. For the spectral clustering algorithm, the \( \alpha \) value used in Equation 6.11 is 15.

To measure overall performance, the purity measure used in speaker clustering literature is adapted to the current context [2, 77]. Dual purity measures are used to evaluate how well closely the automatic clusters match the ‘ground-truth’ sub-arrays, and vice versa. First, define

- \( N_s \): total number of true sub-arrays.
- \( N_c \): total number of found clusters.
- \( N \): total number of microphones.
- \( n_{ij} \): total microphones in cluster \( i \) that are from sub-array \( j \).
- \( n_i \): total microphones in cluster \( i \).
- \( n_{.j} \): total microphones in sub-array cluster \( j \).

The purity of a cluster \( p_i \) is defined as:

\[
p_i = \sum_{j=1}^{N_s} \frac{n_{ij}^2}{n_i^2}
\]

and the average cluster purity \( acp \) is:

\[
acp = \frac{1}{N} \sum_{i=1}^{N_c} p_i \cdot n_i
\]

Similarly, the sub-array purity \( p_j \) and \( asp \) are defined as

\[
p_j = \sum_{i=1}^{N_c} \frac{n_{ij}^2}{n_{.j}^2}
\]

and

\[
asp = \frac{1}{N} \sum_{j=1}^{N_s} p_j \cdot n_{.j}
\]

The \( asp \) gives a measure of how well a sub-array matches only one cluster, and the \( acp \) gives a complementary measure of how well a cluster matches only one sub-array. These scores can be combined to obtain an overall score, \( K = \sqrt{acp \times asp} \). Table 6.1 presents the average score results for \( acp \), \( asp \), and \( K \) for 20 different
6.7. Experimental Results

Table 6.1: Clustering results in terms of average $\bar{a}_{cpl}$, $\bar{a}_{asp}$ and $\bar{K}$ over 20 ad-hoc geometries.

<table>
<thead>
<tr>
<th>Clustering technique</th>
<th>$a_{cpl}$</th>
<th>$a_{asp}$</th>
<th>$\bar{K}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-based clustering</td>
<td>0.855</td>
<td>0.945</td>
<td>0.887</td>
</tr>
<tr>
<td>Spectral clustering</td>
<td>0.867</td>
<td>0.968</td>
<td>0.906</td>
</tr>
</tbody>
</table>

ad-hoc geometries recorded in Data Set A for both the rule-based and spectral clustering algorithms described in Section 6.3.1. Four illustrative examples are plotted in Figure 6.5.

6.7.2 Cluster Proximity Ranking

For the subsequent speech enhancement and recognition evaluation, the rule-based clustering was used to obtain microphone clusters for Data Sets B, C and D. While the spectral showed slightly better performance in Table 6.1, the clusters for the enhancement and recognition experiments are well separated and both methods gave the same clusters.

Following rule-based clustering, each cluster was ranked for its proximity to the speaker by means of the TDOA-based algorithm described in Section 6.4. Table 6.2 details the automatically-determined cluster configurations for the data sets used in the subsequent enhancement and recognition experiments.

Table 6.2: Microphone membership and the average proximity score $\bar{D}_k$ (in ms) to clusters for the ad-hoc array setup in Data Sets B, C and D.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mics.</td>
<td>$\bar{D}_k$</td>
<td>Mics.</td>
</tr>
<tr>
<td>B, D</td>
<td>S1</td>
<td>5,6,7,8</td>
<td>0</td>
<td>1,2,3,4</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>5,6,7,8</td>
<td>0</td>
<td>1,2,3,4</td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>1,2,3,4</td>
<td>0</td>
<td>5,6,7,8</td>
</tr>
<tr>
<td>C</td>
<td>S1</td>
<td>4,5,6</td>
<td>0</td>
<td>7,8</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>4,5,6</td>
<td>0</td>
<td>7,8</td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>1,2,3</td>
<td>0</td>
<td>4,5,6</td>
</tr>
</tbody>
</table>
Figure 6.5: Cluster assignments on the ground truth positions for four geometries, with one geometry per row. In each case, (A) shows the sub-arrays obtained from known microphone positions while (B) and (C) shows the result of rule-based clustering and spectral clustering from the measured noise coherence. The centre microphone in a cluster is shown by encircling line for rule-based clustering.
6.7.3 Blind Beamforming Performance Evaluation

The audio quality of ad-hoc microphone array beamforming as the result of the clustered algorithm is assessed for the purpose of speech enhancement and recognition. The details of the two evaluation procedures are presented below.

**Speech Enhancement**

The conventional method to measure the noise reduction is to compute the amount of speech energy over the noise energy as signal-to-noise ratio after the enhancement. In practice, calculating the true SNR is rarely possible as the speech and noise cannot be separated at the output. Therefore in this chapter, the average segmental signal-plus-noise to noise ratio is computed [67].

The noise level during speech non-active period is first estimated as the average segmental energy as:

\[
\hat{E}_{\text{noise}} = \frac{1}{N_n} \sum_{n=1}^{N_n} \sum_{k=1}^{K} |s(n, k)|^2
\]  

(6.20)

where \(N_n\) is the number of noise-only frames, \(K\) is the frame size, and \(s(n, k)\) is the signal at sample \(k\) in the \(n^{th}\) noise-only frames. The frame SNR (in dB) is then calculated by dividing the segmental energy in the speech-only frame by the average noise energy:

\[
\hat{\text{SNR}}(m) = 10 \log \left( \frac{\sum_{k=1}^{K} |s(m, k)|^2}{\hat{E}_{\text{noise}}} \right)
\]  

(6.21)

where \(m\) is the speech frame number. The average segmental SNR which is effectively a measure of the signal-plus-noise to noise ratio is calculated as:

\[
\overline{\text{SNR}} = \frac{1}{N_s} \sum_{m=1}^{N_s} \hat{\text{SNR}}(m)
\]  

(6.22)

where there are \(N_s\) frames of speech. To obtain the speech- and noise-only portions for frame-based segmental SNR calculation, the phonetic time-aligned transcriptions from the original database where the utterances were taken are used for segmentation.

In addition to calculating segmental SNR to evaluate the noise suppression capability, the overall quality of enhanced speech signal is also evaluated using
Table 6.3: Segmental SNR (dB) of ad-hoc array cluster beamforming in Data Set B. Results are averaged over 30 utterances. The best result for each scenario is highlighted.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Closest cluster</td>
</tr>
<tr>
<td>S1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N2</td>
<td>7.46</td>
<td>8.83</td>
<td>6.43</td>
</tr>
<tr>
<td>N3</td>
<td>10.35</td>
<td>11.28</td>
<td>6.57</td>
</tr>
<tr>
<td>-</td>
<td>12.08</td>
<td>13.42</td>
<td>9.09</td>
</tr>
<tr>
<td>S2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N1</td>
<td>9.89</td>
<td>11.54</td>
<td>4.28</td>
</tr>
<tr>
<td>N2</td>
<td>8.56</td>
<td>11.12</td>
<td>5.38</td>
</tr>
<tr>
<td>N3</td>
<td>10.49</td>
<td>12.38</td>
<td>5.67</td>
</tr>
<tr>
<td>-</td>
<td>12.61</td>
<td>13.25</td>
<td>7.46</td>
</tr>
<tr>
<td>S3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N1</td>
<td>7.49</td>
<td>7.72</td>
<td>6.12</td>
</tr>
<tr>
<td>N2</td>
<td>8.50</td>
<td>8.73</td>
<td>5.44</td>
</tr>
<tr>
<td>N3</td>
<td>10.18</td>
<td>11.06</td>
<td>9.37</td>
</tr>
<tr>
<td>-</td>
<td>10.50</td>
<td>11.09</td>
<td>8.62</td>
</tr>
</tbody>
</table>

The Perceptual Evaluation of Speech Quality (PESQ) measure. It has been shown that the PESQ measure is well correlated with the subjective listening test compared to the segmental SNR since it considers distortions and artifacts introduced in the processing of speech signals by the speech enhancement algorithms [34, 58]. For experiments in this paper, the PESQ software\(^1\) was used to predict the Mean Opinion Score (MOS) of the enhanced speech.

The output segmental SNRs and PESQ measures of the blind beamforming for the closest cluster, the second closest cluster, and using all microphones are presented in Table 6.3 and Table 6.4 for Data Set B respectively and Table 6.5 and Table 6.6 for Data Set C respectively. Because the different blind beamforming methods exhibited very similar performance in terms of these enhancement quality measures, only the Blind MVDR results are shown to simplify presentation and analysis. For comparison, the closest single input channel measure is presented in the same tables.

\(^1\)[Online]. Available: http://www.utdallas.edu/~loizou/speech/software.htm
### Table 6.4: PESQ measures of ad-hoc array cluster beamforming in Data Set B. Results are averaged over 30 utterances. The best result for each scenario is highlighted.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Closest cluster</td>
<td>Second closest cluster</td>
</tr>
<tr>
<td>S1</td>
<td>N1</td>
<td>1.56</td>
<td><strong>1.79</strong></td>
<td>1.30</td>
</tr>
<tr>
<td></td>
<td>N2</td>
<td>1.49</td>
<td><strong>1.71</strong></td>
<td>1.25</td>
</tr>
<tr>
<td></td>
<td>N3</td>
<td>1.64</td>
<td><strong>1.85</strong></td>
<td>1.28</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>1.72</td>
<td><strong>1.99</strong></td>
<td>1.35</td>
</tr>
<tr>
<td>S2</td>
<td>N1</td>
<td>1.46</td>
<td><strong>1.77</strong></td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>N2</td>
<td>1.52</td>
<td><strong>1.71</strong></td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>N3</td>
<td>1.58</td>
<td><strong>1.78</strong></td>
<td>1.09</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>1.58</td>
<td><strong>1.91</strong></td>
<td>1.22</td>
</tr>
<tr>
<td>S3</td>
<td>N1</td>
<td>1.32</td>
<td>1.59</td>
<td>1.31</td>
</tr>
<tr>
<td></td>
<td>N2</td>
<td>1.33</td>
<td>1.58</td>
<td>1.24</td>
</tr>
<tr>
<td></td>
<td>N3</td>
<td>1.40</td>
<td>1.63</td>
<td>1.38</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>1.42</td>
<td>1.71</td>
<td>1.44</td>
</tr>
</tbody>
</table>

### Table 6.5: Segmental SNR (dB) of ad-hoc array cluster beamforming in data set C for closest cluster and second closest cluster. Results are averaged over 30 utterances. The best result for each scenario is highlighted.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Closest cluster</td>
<td>Weighted cluster combination</td>
</tr>
<tr>
<td>S1</td>
<td>N1</td>
<td>9.54</td>
<td>10.85</td>
<td><strong>10.89</strong></td>
</tr>
<tr>
<td></td>
<td>N2</td>
<td>8.64</td>
<td>9.55</td>
<td><strong>10.71</strong></td>
</tr>
<tr>
<td></td>
<td>N3</td>
<td>11.71</td>
<td>12.20</td>
<td><strong>12.22</strong></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>12.90</td>
<td>13.67</td>
<td><strong>14.00</strong></td>
</tr>
<tr>
<td>S2</td>
<td>N1</td>
<td>8.73</td>
<td>9.51</td>
<td><strong>10.20</strong></td>
</tr>
<tr>
<td></td>
<td>N2</td>
<td>7.40</td>
<td>7.79</td>
<td>8.44</td>
</tr>
<tr>
<td></td>
<td>N3</td>
<td>10.53</td>
<td><strong>11.45</strong></td>
<td>11.35</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>11.48</td>
<td>11.64</td>
<td><strong>11.97</strong></td>
</tr>
<tr>
<td>S3</td>
<td>N1</td>
<td>6.55</td>
<td>6.37</td>
<td>7.36</td>
</tr>
<tr>
<td></td>
<td>N2</td>
<td>6.43</td>
<td>5.59</td>
<td>7.02</td>
</tr>
<tr>
<td></td>
<td>N3</td>
<td>8.80</td>
<td>8.49</td>
<td>9.95</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>10.26</td>
<td>10.44</td>
<td>10.43</td>
</tr>
</tbody>
</table>
Table 6.6: PESQ measures of ad-hoc array cluster beamforming in data set C for closest cluster and second closest cluster. Results are averaged over 30 utterances. The best result for each scenario is highlighted.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Closest cluster</td>
</tr>
<tr>
<td>S1</td>
<td>N1</td>
<td>1.66</td>
<td>1.83</td>
</tr>
<tr>
<td></td>
<td>N2</td>
<td>1.63</td>
<td><strong>1.91</strong></td>
</tr>
<tr>
<td></td>
<td>N3</td>
<td>1.76</td>
<td><strong>2.01</strong></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>1.86</td>
<td><strong>2.16</strong></td>
</tr>
<tr>
<td>S2</td>
<td>N1</td>
<td>1.28</td>
<td>1.42</td>
</tr>
<tr>
<td></td>
<td>N2</td>
<td>1.30</td>
<td>1.34</td>
</tr>
<tr>
<td></td>
<td>N3</td>
<td>1.37</td>
<td>1.53</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>1.35</td>
<td>1.52</td>
</tr>
<tr>
<td>S3</td>
<td>N1</td>
<td>1.10</td>
<td>1.22</td>
</tr>
<tr>
<td></td>
<td>N2</td>
<td>1.26</td>
<td>1.36</td>
</tr>
<tr>
<td></td>
<td>N3</td>
<td>1.27</td>
<td>1.37</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>1.39</td>
<td>1.48</td>
</tr>
</tbody>
</table>

**Speech Recognition**

While improvement in SNR gives a good indication of performance of the enhanced signal in terms of noise reduction, it is appropriate to confirm the proposed clustered beamforming approach when it is used as a front-end for automatic speech recognition.

Experiments were conducted on Data Set D using the WebASR web service\(^2\), which was configured to run a system based on the AMI MDM system [29, 30]. The WebASR system is a state-of-the-art large vocabulary system, optimised for conversational speech such as in meetings, from international English speakers, and trained using table-top microphones. As a benchmark comparison, speech recognition results generated from the original clean database files gave a WER of 26.1%. As it is not optimised specifically for read speech from the WSJ corpus, this baseline performance is lower than might be achieved with a task-specific ASR system, however the WebASR system was chosen as it provides robust results when using speech re-recorded on distant microphones. WebASR also makes it easy for other researchers to benchmark array processing methods using

\(^2\)[Online]. Available at http://www.webasr.org/
6.7. Experimental Results

Table 6.7: % word error rate from various blind beamforming methods on Data Set D. For comparison, WER of 26.1% is achieved using the original clean corpus recordings on the same ASR system.

<table>
<thead>
<tr>
<th>Spk. Pos.</th>
<th>Technique</th>
<th>Closest clust.</th>
<th>2nd closest clust.</th>
<th>All microphones</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Single Ch.</td>
<td>43.2</td>
<td>64.0</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Delay-sum</td>
<td>35.7</td>
<td>56.5</td>
<td>44.4</td>
</tr>
<tr>
<td></td>
<td>MVDR</td>
<td>36.0</td>
<td>59.8</td>
<td>48.1</td>
</tr>
<tr>
<td></td>
<td>Superdirective</td>
<td>34.8</td>
<td>61.4</td>
<td>55.5</td>
</tr>
<tr>
<td>S2</td>
<td>Single Ch.</td>
<td>45.7</td>
<td>69.8</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Delay-sum</td>
<td>36.8</td>
<td>67.8</td>
<td>49.6</td>
</tr>
<tr>
<td></td>
<td>MVDR</td>
<td>36.2</td>
<td>76.4</td>
<td>54.8</td>
</tr>
<tr>
<td></td>
<td>Superdirective</td>
<td>35.5</td>
<td>83.3</td>
<td>64.3</td>
</tr>
<tr>
<td>S3</td>
<td>Single Ch.</td>
<td>53.4</td>
<td>62.6</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Delay-sum</td>
<td>46.3</td>
<td>56.5</td>
<td>49.9</td>
</tr>
<tr>
<td></td>
<td>MVDR</td>
<td>45.8</td>
<td>64.2</td>
<td>54.4</td>
</tr>
<tr>
<td></td>
<td>Superdirective</td>
<td>46.0</td>
<td>67.4</td>
<td>61.2</td>
</tr>
</tbody>
</table>

the same ASR system used in this paper. The speech recognition results are presented in Table 6.7 for single channel input and the outputs from the blind Delay-Sum, MVDR, and Superdirective techniques.

6.7.4 Experiments on Weighted Cluster Combination

To investigate whether performance may be improved by combining cluster beamformer outputs, a first set of experiments was conducted on Data Set B to study the effect of cluster weight. The weighted cluster combination beamforming method described in Section 6.5.2 was used to combine the two clusters with varying cluster weights. Note that only the blind Delay-Sum beamforming method was used for these experiments (hence results vary from those reported using MVDR on the same data in Table 6.3). As there are only two clusters, the weighting parameters $\alpha_k$ from Equation 6.15 were set to $\alpha_1$ and $\alpha_2 = 1 - \alpha_1$, with $\alpha_1$ varied from 0 to 1.

The weighted cluster combination beamforming is evaluated in terms of the
Table 6.8: % word error rate of weighted cluster delay-sum combination.

<table>
<thead>
<tr>
<th>Speaker Position</th>
<th>Weighted Cluster Comb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>35.5</td>
</tr>
<tr>
<td>S2</td>
<td>36.2</td>
</tr>
<tr>
<td>S3</td>
<td>42.3</td>
</tr>
</tbody>
</table>

SNR improvement, determined by subtracting the input SNR of the closest microphone from the SNR of the weighted cluster combination output. Figure 6.6 shows the SNR improvement of the weighted cluster combination for Data Set B in different speaker positions for the varying weights. The SNR improvement obtained from the closest cluster and using all microphones are also plotted for comparison.

These preliminary results suggest two effects. First, there seems little benefit to combining clusters when one of the clusters is considerably closer to the speaker and further from other noises than the others (i.e. for S1 and S2. See Figure 6.4 and the second cluster proximity measure in Table 6.2). Second, when it is worth combining multiple clusters due to their similar distance to the speaker (i.e. for S3), the weighting of each depends on their relative size (number of microphones), as suggested by the theoretical motivations in Section 6.2. The plot of SNR improvement versus cluster weight in Figure 6.6(c) shows that even though the two clusters independently offer different SNRs due to their positioning with respect to speech and noise sources, it is still optimal to equally weight the two clusters, as they have the same size and are at a similar distance to the speaker.

Following these observations, a second set of experiments was conducted on Data Set C. For this dataset, the clusters are less separated and so there is generally more motivation for their combination. Tables 6.5 and 6.6 give the SNR and PESQ results obtained from a weighted combination of the first two clusters, where the weight was set to reflect their relative sizes: that is, weights of (0.6,0.4), (0.6,0.4) and (0.5, 0.5) were used for speaker positions S1, S2 and S3, respectively.
6.7. Experimental Results

Figure 6.6: The delay-sum SNR improvement of weighted cluster combination for noise condition N1 in different speaker positions by varying weight $\alpha_1$ for the closest cluster and weight $\alpha_2$ for the second closest cluster.
6.8 Discussion

6.8.1 Clustering of Microphones

The high purity measures in Table 6.1 indicate that both clustering algorithms using the magnitude squared coherence feature well-approximate the sub-arrays formed from known microphone positions. When the separation between clusters is clear, as in Figure 6.5(i), the algorithm succeeds.

The lower value of $\bar{a}cp$ compared to $\bar{a}sp$ in Table 6.1 for recorded geometries shows that the automatic measure tends to create larger clusters for microphone separations near the threshold value, indicating that the measured coherence tends to exceed that predicted by the diffuse model in this particular environment. Examples of this occurring are shown in Figure 6.5(ii)-(iii). For Figure 6.5(iii), however, the spectral clustering separates them as two distinct clusters since elements in each are tightly spaced making the cut between clusters more obvious.

The array geometry in Figure 6.5(iv) presents a case when all microphones are somewhat continuously close to each other. Assuming known microphone positions, the algorithm merges all microphones together to form a single cluster. When the noise signal is used, the rule-based algorithm forms 2 distinct clusters as the measured coherence between microphone 5 and 6 is below the threshold in this particular situation. For spectral clustering however, the clusters are overlapping and it is very difficult to determine the number of clusters since the eigenvalues are likely to have continuous values with no well-defined gap. Therefore, the number of clusters is set to be one. Such a situation shows that perhaps more sophisticated methods may benefit from incorporating multiple features into the clustering decision, for instance to encode structural regularity as well as just inter-element proximity.

6.8.2 Blind Beamforming Experiments

Experiments on Data Set B provide an example situation where the speaker is relatively close to one of the clusters and far from the other. In this situation,
6.8. Discussion

Figure 6.7: Box plot of the TDOA values applied to closest cluster, second closest cluster, and all microphones for speaker at position S1 with noise condition N1 from Data Set B. The attributes of plots are similar to box plot in Figure 6.1.
considering the results in Table 6.3 and 6.4, using the closest cluster for beamforming gives the best performance in terms of SNR and PESQ compared to the closest channel or the full array. For the speaker at position S3 however, the speaker is roughly the same distance from both clusters. This is reflected in the lower proximity score $D_k$ in Table 6.2 for the second closest cluster in the S3 case. In this case, the SNR and PESQ measures show similar performance between the closest cluster and the full array, with the full array offering higher PESQ.

For the second closest cluster, the beamforming techniques do not show improvement over the single channel. This is likely attributed to the minor steering errors which effectively result in signal degradation, as well as the lower input SNR. Depending on the location of the speaker, the beamformer output from the full array may not give improvement over the output of closest cluster with less number of microphones. While using more microphones in theory increases the overall array gain, in this practical instance the degradation is attributable to the large inter-microphone spacings leading to erroneous blind TDOA estimation. To phase-align the microphone signals, the TDOA is computed by selecting a reference microphone and calculating delays relative to this reference. Unfortunately the calculation of delays in this way causes inaccuracies between more distant pairs.

The cross-correlation function between two microphones which are spatially close will be dominated by a peak corresponding to the TDOA difference, as they receive signals which have otherwise undergone very similar acoustic transfer functions from the source. For microphones with large distances however, the two impulse responses are likely to be different, increasing the probability of reflections obscuring the cross-correlation peak. To illustrate this phenomenon, the TDOA values computed from the speech frames between microphones in the closest cluster, second closest cluster, and using all microphones are presented in the form of box plot in Figure 6.7.

Figure 6.7 illustrates that delays calculated from microphones in the closest cluster are reasonably accurate, in which the mean, median, and expected ground truth delays have consistently similar values. Similar delay accuracy is observed
6.8. Discussion

in the second closest cluster but with a larger standard deviation. Delays calculated between distant microphone pairs, however, shows large inconsistencies (e.g. TDOA$_{51}$, TDOA$_{61}$, TDOA$_{71}$, and TDOA$_{81}$) due to the large distance between the closest microphone as the reference microphone and the 5$^{th}$, 6$^{th}$, 7$^{th}$, 8$^{th}$ closest microphone to the speaker respectively.

In the second set of experiments conducted on Data Set C, results in Table 6.5 and 6.6 show that in general the weighted cluster combination of the two closest clusters improves speech quality compared to either the closest cluster or the full array. For a speaker at position S2 and S3, the extremely low proximity score of the second closest cluster of 0.5 and 0.2 in Table 6.2 suggests that the distance from the speaker to the two closest clusters is approximately the same. In this situation, it is beneficial to combine multiple clusters rather than simply select the closest one. This shows that the proximity score $D_k$ may be used to indicate whether it is worth combining clusters based on their relative distance to the speaker.

6.8.3 Speech Recognition Experiments

The speech recognition performance from the closest cluster beamforming output shows significant improvement over a single channel for all investigated beamforming techniques. Delay-sum, MVDR, and Superdirective show similar improvement albeit slightly higher for the last two techniques. Position S3 shows the smallest improvement of 7.1% compared to the single channel input, with WER of 53.4% compared to position S1 which improves 7.5% from 43.2%, and position S2 which improves 8.9% from 45.7% for delay-sum. For this position, the lower single channel WER and its beamforming output in the closest cluster is attributed to the lower input SNR and PESQ measures (Table 6.3 and 6.4) as the speaker is not facing the microphones directly.

In the second closest cluster, while delay-sum shows improvement compared to the single channel, this is not the case for MVDR and superdirective beamforming. This degradation of performance is likely attributed to the steering error due to inaccuracy in direct delay estimations, as discussed above. This is more evident in the MVDR and superdirective as they are more sensitive to steering errors.
than delay-sum \cite{81}. For this same reason, together with the fact that some of the more distant microphones have lower speech quality, beamforming using the full array also degrades the performance.

As explained in Section 2.8, to reduce the sensitivity of MVDR and superdirective to this type of steering error, a small scalar is usually added to the diagonal of the coherence matrix. The scalar $\mu$ depends on the choice of constraint $T_{se}$ in Equation 2.53. Decreasing the sensitivity will result in a large value of $\mu$, biasing the effective noise matrix towards the identity matrix as in delay-sum: as $\mu$ approaches infinity in Equation 2.62, the performance of MVDR and superdirective is therefore expected to be equal to the delay-sum. While MVDR and superdirective prove to be beneficial when using a cluster of spatially close microphones due to their optimised noise coherence suppression, in the case of blind estimation of delays, the delay-sum shows to be a more robust beamforming technique.

### 6.8.4 Weighted Cluster Combination Experiments

Experiments on the weighted cluster combination beamforming show that, depending on the relative positions of speakers, microphones and noises, optimal performance may be obtained when the contributions of cluster are taken into account. To give more insight into cluster combination, Figure 6.6(a) presents a plot of SNR with the peak result occurring when $\alpha_1$ and $\alpha_2$ are set to be 0.8 and 0.2 respectively. Similarly in Figure 6.6(c), the optimal performance for speaker position S3 is obtained when $\alpha_1$ and $\alpha_2$ are set to be equal 0.5. For speaker at position S2, however, using the closest cluster by itself yields the optimal performance. The best performance of weighted cluster combination for speakers at position S1 and S3 are only slightly higher than their closest cluster beamforming performance. This is likely attributed to the large distance that separates both clusters, causing small contributions of the second closest cluster to the overall performance. For position S1 in particular, only a small contribution is expected due to the fact that the second closest cluster is located close to the noise source. Using all microphones directly without any clustering results in negative SNR improvement - this is somewhat counter-intuitive, as it may be expected that this would correspond to the performance of the equally weighted cluster combination.
6.9. Conclusion

(e.g. $\alpha_1 = \alpha_2 = 0.5$). The difference is again due to the inaccuracy in TDOA estimation for more distant microphone pairs, as explained above - in the case of weighted cluster combination, only the single inter-cluster delay is calculated over a longer distance, while for the unclustered approach, many pairs will involve relatively large distances.

To confirm the effect of cluster combination on speech recognition performance, the delay-sum weighted cluster combination beamformer for S3 was tested on Data Set D, gave 42.3% WER, compared with 46.3% using the closest cluster and 49.9% using all microphones.

These results, as well as theoretical motivations, indicate that optimal cluster weights depend on cluster size and relative proximity. The effect of combining clusters with weights according to cluster size on Data Set C was discussed in Section 6.8.2 above.

6.9 Conclusion

In this chapter, a novel clustered approach to blind beamforming from ad-hoc microphone arrays has been proposed. For the first step, two microphone clustering algorithms were proposed to group microphones using only the knowledge of noise coherence. In a second step, the clusters were ranked based on their proximity to the speaker using TDOA information. Finally, two methods for achieving microphone array beamforming using these clusters were investigated in the context of speech enhancement and recognition: closest cluster beamforming, and weighted cluster combination beamforming. Experiments were conducted on a new database recorded for this purpose in a typical meeting room environment. Eight microphones were used in a variety of placements to simulate ad-hoc arrangements on a meeting table. Experiments validated the partitioning provided by the clustering algorithms, as well as the effect of subsequent beamforming on the SNR, PESQ, and the word error rate.

Depending on the relative distance from the cluster of microphones to the speaker as indicated by their proximity score, using a cluster or multiple clusters
in the same time can provide better performance than a larger array. An underlying cause of this improvement is the fact that larger inter-microphone distances can lead to erroneous delay estimation for blind steering vector formulations. The speech recognition experiments further confirm the benefit of clustering, as both clustered beamforming methods (closest cluster and weighted cluster combination) show significant improvement over both the single channel input and the full set of microphones.

Amongst the blind beamforming techniques investigated, it was found that delay-sum beamforming is the most robust when using the full set of microphones, as it is less sensitive to steering errors than MVDR and superdirective beamforming. When only the closest cluster is used, however, MVDR and superdirective improve more relative to delay-sum, as these methods are optimal for reducing the ambient noise when more accurate steering vectors are used. Overall, however, all beamforming methods offer similar speech recognition performance from the clustered beamforming, and therefore delay-sum is a good candidate in practical deployments due to its simplicity and decreased sensitivity to delay estimation errors. While not investigated in this paper, it is however noted that MVDR and superdirective beamforming provide a potential means of trading off robustness to steering errors with noise reduction by varying the white noise gain constraint.

Experiments on the weighted cluster combination beamforming indicate that combining microphones at the cluster level, rather than individual level, offers improved robustness for blind beamforming when the clusters are similarly proximate to the speaker. The proposed cluster proximity measure offers a promising means of detecting when this cluster combination should be used.

Based on the above findings, further research into beamforming from ad-hoc microphone arrays will investigate other measures for ranking clusters, and automatically determining cluster weights for combination. In true ad-hoc situations, characteristics other than just distance to the speaker should be considered, as microphones may be of widely varying quality and response. Finally, it is noted that while the research in this thesis has been constrained to propose solutions for unknown geometries, in true ad-hoc scenarios robust methods must also consider potential differences in microphone response and synchronisation.
6.9. Conclusion

The research contained in this chapter is the subject of the following fully-reviewed publications:


Chapter 7

Conclusions and Future Research

7.1 Summary of Contributions

This dissertation has investigated microphone-array-based speech recognition from ad-hoc array geometry in meetings. In ad-hoc situations when the microphones have not been systematically positioned, microphone and likely source locations are usually not known and beamforming must be achieved blindly. The work in this thesis has concentrated on researching and developing methods to beamforming from ad-hoc microphone arrays. Within this broad problem, the following specific goals were set as the main objectives of this thesis:

(i) To investigate the impact of using ad-hoc microphone arrays on speech recognition performance.

(ii) To propose novel approaches to beamforming from ad-hoc microphone arrays when microphone positions and likely source locations are not known.

(iii) To consider the implementation of the proposed algorithms within the constraints of meeting room environments.

and the specific research objectives were to:

(i) To review and identify limitations in the current state of research into ad-hoc microphone arrays for speech recognition applications.
(ii) To quantify the word accuracy performance of ASR system when there is error in microphone placements from the original assumed positions.

(iii) To propose two general approaches to blindly estimate the steering vector for beamforming when microphone positions and likely source locations are not known.

(iv) To propose a novel approach to blind speech separation using microphone array in multi-speaker environments.

(v) To propose novel techniques in dealing with randomly placed microphones and speaker positions.

(vi) To assess each of the proposed techniques and report their performance based on speech recognition accuracy using recorded data from the meeting room environments.

The major original contributions resulting from this work has achieved each of these aims and objectives and summarised as follows:

(i) Chapter 3 presented a review of microphone-array-based speech recognition followed by beamforming from ad-hoc microphone arrays in the meeting task. The research directions were proposed motivated by identifying the limitations in the literature. Subsequently, comparison of robust beamforming approaches in terms of speech recognition accuracy were presented with a simulated placement errors. It was found that the investigated techniques were robust to certain degree of errors.

(ii) Proposed two general approaches to beamforming from ad-hoc microphone arrays when microphone positions and likely source locations are not known. The automatic calibration approach first determine the unknown microphone positions through array calibration methods and then use the traditional geometrical formulation for the steering vector, while blind estimation approach directly estimate the steering vector without regard to the microphone and source locations. The evaluation between the two approaches
for blind beamforming from small ad-hoc arrays showed comparable performance between the two approaches.

(iii) Analysis of a novel array shape calibration technique. The technique is based on the assumed diffuse noise model of the background noise and does not require purpose-built devices or known calibration signals. The calibrated array geometry showed accurate geometry with small errors with little effect on the beamformer’s look direction.

(iv) Proposed blind source separation using microphone arrays for the problem of overlapped speech in meetings. The proposed algorithm were tested on overlapped speech database with a performance close to the close-talking microphone and significant increase over a single distant microphone.

(v) Proposed a novel clustered approach for beamforming from ad-hoc microphone arrays. Overall, speech enhancement results showed that using a cluster of microphones located close to the speaker can provide better performance than a larger array. An underlying cause of this improvement is the fact that large inter-microphone distances can lead to an erroneous delay estimation for blind steering vector formulations. The speech recognition experiments further confirm the benefit of clustering, as both clustered beamforming methods (closest cluster and weighted cluster combination) showed significant improvement over both the single channel input and the full set of microphones.

(vi) Studied the performance of delay-sum, MVDR, and superdirective beamforming in the real ad-hoc arrays scenarios. Results showed that amongst the blind beamforming techniques investigated, it was found that delay-sum beamforming is the most robust when using the full set of microphones, as it is less sensitive to steering errors than MVDR and superdirective beamforming. When only the closest cluster is used, however, MVDR and superdirective improve more relative to delay-sum, as these methods are optimal for reducing the ambient noise when more accurate steering vectors are used. Overall, however, all beamforming methods offer similar noise reduction
and speech recognition performance from the clustered beamforming, and therefore delay-sum is a good candidate in practical deployments due to its simplicity and decreased sensitivity to delay estimation errors.

### 7.2 Additional Contributions

In addition to the main contributions of the thesis above, the following contributions were made that were not included in this thesis. While they are not directly related to the principle topic of this thesis, they are still within the broad topic of microphone-array-based speech recognition.

1. A posterior approach for microphone array based speech recognition. In this work, instead of enhancing speech signals, the posterior phone probabilities were enhanced which are then used in a tandem Artificial Neural Network - Hidden Markov Model (ANN-HMM). The new approach directly enhances posterior probabilities, so might be called posterior beamforming, in the sense of channel selection and accumulation. The approach is an intermediate combination approach, somewhere between acoustic beamforming and hypothesis integration (e.g., systems that use Recogniser Output Voting Error Reduction (ROVER)). Posterior beamforming has several potential advantages: (1) time-delay estimation is not required, assuming the posteriors change smoothly; (2) heterogeneous channels can be combined easily; (3) channel selection is more meaningful based on the posteriors.


2. FPGA Implementation of Dual-Microphone Delay-and-Sum Beamforming for In-Car Speech Enhancement and Recognition. In this work, FPGA design of a dual microphone delay-and-sum beamformer for speech enhancement specifically designed toward cheaper and efficient solutions for in-car environments was presented. Experimental results show that the proposed
design can produce output waveforms close to those generated by a theoretical (floating-point) model with modest usage of FPGA resources. These results were confirmed with comparable speech recognition performance between the proposed design and those generated by theoretical (floating-point) model.


7.3 Future Research

In this thesis, solutions towards the problem of speech recognition using ad-hoc microphone arrays in the case of unknown microphone and speaker locations were investigated, with results presented for medium-vocabulary speech recognition in the meeting task. To continue on the research presented in this thesis, some new directions for future research are listed below:

• While the research in this thesis has been constrained to propose solutions for unknown geometries of ad-hoc microphone arrays, in true ad-hoc scenarios robust methods must also consider potential differences in microphone response and synchronisation.

• The automatic calibration technique has been proposed as a solution to the problem of overlapped speech in meetings. Unfortunately, the automatic calibration has constraint such as when the array is large with wide inter-microphone spacing. Computational cost is also an issue to do speaker localisation over a large meeting area. A useful avenue for future research would be to investigate the use of TDOA delays to simultaneously beamform to the active speakers without calibration and localisation steps.

• In the clustered beamforming approach proposed in Chapter 6, experiments on the weighted cluster combination beamforming indicate that combining
microphones at the cluster level, rather than individual level, offers improved robustness for blind beamforming. It would be desirable to automatically determining cluster weights for combination in order to obtain optimal speech recognition accuracy.

- The ad-hoc array configurations in this thesis were necessarily restricted with maximum number of 8-elements due to the lack of appropriate microphone array hardware. It would be interesting to try the proposed clustered algorithm with various ad-hoc array configurations of different number and size of clusters, as well as the configuration of speech and noise sources.

- Speech recognition evaluation of the proposed clustered approach in the NIST meeting data.

- The microphone-array-based speech recognition in this thesis followed the conventional approach in which the array processing acts as a front-end pre-processing to generate cleaner signal for the input of speech recognition system. It would be an interesting research topic to integrate the blind beamforming approach and the speech recognition system for the aim of maximising speech recognition performance.
Bibliography


[44] H. K. Maganti, D. Gatica-Perez, and I. McCowan, “Speech enhancement and


[61] ———, “Array shape calibration using sources in unknown locations—part ii: Near-field sources and estimator implementation,” IEEE Transactions on


[78] K. V. Stavropoulos and M. A., “Array calibration in the presence of unknown


Appendix A

Supplemental Mathematical Derivations

A.1 Minimum Variance Distortionless Response

The constrained optimisation problem may be stated as follows:

Minimise

\[
\min_w w^H \Phi_{\nu \nu} w
\]  \hspace{1cm} (A.1)

subject to distortionless constraint given by

\[
w^H d = 1
\]  \hspace{1cm} (A.2)

To solve the constrained optimisation problem, the method of Lagrange multipliers is used. First define the function need to be minimised:

\[
J = w^H \Phi_{\nu \nu} w + \Re \left[ \lambda^* w^H d - 1 \right]
\]  \hspace{1cm} (A.3)

where \( \lambda \) is a complex Lagrange multiplier. Taking the complex gradient with respect to \( w \) gives

\[
\frac{\partial}{\partial w^*} (w^H w) + \frac{\partial}{\partial w^*} \left( \Re \left[ \lambda^* w^H d - 1 \right] \right) = 0
\]  \hspace{1cm} (A.4)

Using the rules of differentiation, the first term in the summation

\[
\frac{\partial}{\partial w^*} (w^H \Phi_{\nu \nu} w) = \Phi_{\nu \nu} w
\]  \hspace{1cm} (A.5)
and the second term
\[
\frac{\partial}{\partial w^*}(\mathbb{R} [\lambda^* w^H d - 1]) = \lambda^* d \quad \text{(A.6)}
\]

Substituting the results of A.5 and A.6 into Equation A.4 gives
\[
\Phi_{vv} w + \lambda^* d = 0 \quad \text{(A.7)}
\]

Solving for optimum value \(w_o\),
\[
w_o = -\lambda^* \Phi_{vv}^{-1} d \quad \text{(A.8)}
\]

To obtain solution for \(\lambda\), the Hermitian transpose is taken on the both side of Equation A.8 and then multiply by \(d\) which give
\[
\lambda = -\frac{1}{d^H \Phi_{vv}^{-1} d} \quad \text{(A.9)}
\]

where the constraint \(w^H d = 1\) is used and the fact that \(\Phi_{vv}^{-H} = \Phi_{vv}^{-1}\).

Substituting Equation A.9 into Equation A.8, the desired solution for optimum weight vector is found as
\[
w_o = \frac{\Phi_{vv}^{-1} d}{d^H \Phi_{vv}^{-1} d} \quad \text{(A.10)}
\]